Exponential Families and Kernels
Lecture 1

Alexander J. Smola
Alex.Smola@nicta.com.au

Machine Learning Program
National ICT Australia
RSISE, The Australian National University
Outline

Exponential Families
- Maximum likelihood and Fisher information
- Priors (conjugate and normal)

Conditioning and Feature Spaces
- Conditional distributions and inner products
- Clifford Hammersley Decomposition

Applications
- Classification and novelty detection
- Regression

Applications
- Conditional random fields
- Intractable models and semidefinite approximations
Model

- Log partition function
- Expectations and derivatives
- Maximum entropy formulation

Examples

- Normal distribution
- Discrete events
- Laplacian distribution
- Poisson distribution
- Beta distribution

Estimation

- Maximum Likelihood Estimator
- Fisher Information Matrix and Cramer Rao Theorem
- Normal Priors and Conjugate Priors
The Exponential Family

Definition
A family of probability distributions which satisfy

\[ p(x; \theta) = \exp(\langle \phi(x), \theta \rangle - g(\theta)) \]

Details
- \( \phi(x) \) is called the sufficient statistics of \( x \).
- \( \mathcal{X} \) is the domain out of which \( x \) is drawn (\( x \in \mathcal{X} \)).
- \( g(\theta) \) is the log-partition function and it ensures that the distribution integrates out to 1.

\[ g(\theta) = \log \int_{\mathcal{X}} \exp(\langle \phi(x), \theta \rangle) dx \]
Example: Binomial Distribution

Tossing coins

With probability $p$ we have heads and with probability $1 - p$ we see tails. So we have

$$p(x) = p^x(1 - p)^{1-x} \text{ where } x \in \{0, 1\} =: \mathcal{X}$$

Massaging the math

$$p(x) = \exp \log p(x)$$
$$= \exp (x \log p + (1 - x) \log (1 - p))$$
$$= \exp \left( \langle (x, 1 - x), (\log p, \log (1 - p)) \rangle \right)$$

The Normalization  Once we relax the restriction on $\theta \in \mathbb{R}^2$ we need $g(\theta)$ which yields

$$g(\theta) = \log \left( e^{\theta_1} + e^{\theta_2} \right)$$
Example: Binomial Distribution
Atomic decay

At any time, with probability $\theta dx$ an atom will decay in the time interval $[x, x + dx]$ if it still exists. Consulting your physics book tells us that this gives us the density

$$p(x) = \theta \exp(\theta x) \text{ where } x \in [0, \infty) =: \mathcal{X}$$

Massaging the math

$$p(x) = \exp\left(\langle -x, \theta \rangle - \log \theta\right)$$
Example: Laplace Distribution
Example: Normal Distribution

Engineer’s favorite

\[ p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{1}{2\sigma^2}(x - \mu)^2 \right) \] where \( x \in \mathbb{R} =: \mathcal{X} \)

Massaging the math

\[ p(x) = \exp \left( -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right) \]

\[ = \exp \left( \langle (x, x^2), \theta \rangle - \left( \frac{\mu^2}{2\sigma^2} + \frac{1}{2} \log(2\pi\sigma^2) \right) \right) \]

Finally we need to solve \((\mu, \sigma^2)\) for \(\theta\). Tedious algebra yields \(\theta_2 := -\frac{1}{2}\sigma^{-2}\) and \(\theta_1 := \mu\sigma^{-2}\). We have

\[ g(\theta) = -\frac{1}{4} \theta_1^2 \theta_2^{-1} + \frac{1}{2} \log 2\pi - \frac{1}{2} \log -2\theta_2 \]
Many discrete events
Assume that we have disjoint events \([1..n] =: \mathcal{X}\) which all may occur with a certain probability \(p_x\).

Guessing the answer
Use the map \(\phi : x \rightarrow e_x\), that is, \(e_x\) is an element of the canonical basis \((0, \ldots, 0, 1, 0, \ldots)\). This gives

\[ p(x) = \exp(\langle e_x, \theta \rangle - g(\theta)) \]

where the normalization is

\[ g(\theta) = \log \sum_{i=1}^{n} \exp(\theta_i) \]
Example: Multinomial Distribution
Example: Poisson Distribution

Limit of Binomial distribution

Probability of observing $x \in \mathbb{N}$ events which are all independent (e.g. raindrops per square meter, crimes per day, cancer incidents)

$$p(x) = \exp \left( x \cdot \theta - \log \Gamma(x + 1) - \exp(\theta) \right).$$

Hence $\phi(x) = x$ and $g(\theta) = e^{\theta}$.

Differences

- We have a normalization dependent on $x$ alone, namely $\Gamma(x + 1)$. This leaves the rest of the theory unchanged.
- The domain is countably infinite.
- Effectively this assumes the measure $\frac{1}{x!}$ on the domain $\mathbb{N}$. 
Example: Poisson Distribution
Example: Beta Distribution

Usage
Often used as prior on Binomial distributions (it is a conjugate prior as we will see later).

Mathematical Form

\[ p(x) = \exp\langle (\log x, \log(1 - x)), (\theta_1, \theta_2) \rangle - \log B(\theta_1 + 1, \theta_2 + 1) \]

where the domain is \( x \in [0, 1] \) and

\[ g(\theta) = \log B(\theta_1 + 1, \theta_2 + 1) \]
\[ = \log \Gamma(\theta_1 + 1) + \log \Gamma(\theta_2 + 1) - \log \Gamma(\theta_1 + \theta_2 + 2) \]

Here \( B(\alpha, \beta) \) is the \textit{Beta} function.
Example: Beta Distribution
Example: Gamma Distribution

Usage

- Popular as a prior on coefficients
- Obtained from integral over waiting times in Poisson distribution

Mathematical Form

\[ p(x) = \exp(\langle (\log x, x), (\theta_1, \theta_2) \rangle - \log \Gamma(\theta_1+1) + (\theta_1+1) \log - \theta_2) \]

where the domain is \( x \in [0, \infty] \) and

\[ g(\theta) = \log \Gamma(\theta_1 + 1) + (\theta_1 + 1) \log - \theta_2 \]

Note that \( \theta \in [0, \infty) \times (-\infty, 0) \).
Example: Gamma Distribution

\[ \text{Graph showing different gamma distributions with different } \alpha \text{ values.} \]
<table>
<thead>
<tr>
<th>Name</th>
<th>$\phi(x)$</th>
<th>Domain</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial</td>
<td>$(x, 1 - x)$</td>
<td>${0, 1}$</td>
<td>discrete</td>
</tr>
<tr>
<td>Multinomial</td>
<td>$e^x$</td>
<td>${1, \ldots, n}$</td>
<td>discrete</td>
</tr>
<tr>
<td>Poisson</td>
<td>$x$</td>
<td>$\mathbb{N}_0$</td>
<td>discrete</td>
</tr>
<tr>
<td>Laplace</td>
<td>$x$</td>
<td>$[0, \infty)$</td>
<td>Lebesgue</td>
</tr>
<tr>
<td>Normal</td>
<td>$(x, x^2)$</td>
<td>$\mathbb{R}$</td>
<td>Lebesgue</td>
</tr>
<tr>
<td>Beta</td>
<td>$(\log x, \log(1 - x))$</td>
<td>$[0, 1]$</td>
<td>Lebesgue</td>
</tr>
<tr>
<td>Gamma</td>
<td>$(\log x, x)$</td>
<td>$[0, \infty)$</td>
<td>Lebesgue</td>
</tr>
<tr>
<td>Wishart</td>
<td>$(\log</td>
<td>X</td>
<td>, X)$</td>
</tr>
<tr>
<td>Dirichlet</td>
<td>$\log x$</td>
<td>$x \in \mathbb{R}_+^n, |x|_1 = 1$</td>
<td>Lebesgue</td>
</tr>
</tbody>
</table>
Recall

Definition
A family of probability distributions which satisfy

\[ p(x; \theta) = \exp(\langle \phi(x), \theta \rangle - g(\theta)) \]

Details
- \( \phi(x) \) is called the **sufficient statistics** of \( x \).
- \( X \) is the domain out of which \( x \) is drawn (\( x \in X \)).
- \( g(\theta) \) is the **log-partition function** and it ensures that the distribution integrates out to 1.

\[ g(\theta) = \log \int_X \exp(\langle \phi(x), \theta \rangle) dx \]
$g(\theta)$ generates cumulants:

$$g(\theta) = \log \int \exp(\langle \phi(x), \theta \rangle) dx$$

Taking the derivative wrt. $\theta$ we can see that

$$\partial_\theta g(\theta) = \frac{\int \phi(x) \exp(\langle \phi(x), \theta \rangle) dx}{\int \exp(\langle \phi(x), \theta \rangle) dx} = \mathbb{E}_{x \sim p(x;\theta)} [\phi(x)]$$

$$\partial_\theta^2 g(\theta) = \text{Cov}_{x \sim p(x;\theta)} [\phi(x)]$$

...and so on for higher order cumulants ...

**Corollary:**

$g(\theta)$ is convex
Benefits: Simple Estimation

**Likelihood of a set:** Given $X := \{x_1, \ldots, x_m\}$ we get

$$p(X; \theta) = \prod_{i=1}^{m} p(x_i; \theta) = \exp \left( \sum_{i=1}^{m} \langle \phi(x_i), \theta \rangle - mg(\theta) \right)$$

**Maximum Likelihood**

We want to minimize the negative log-likelihood, i.e.

$$\text{minimize}_{\theta} \ g(\theta) - \left\langle \frac{1}{m} \sum_{i=1}^{m} \phi(x_i), \theta \right\rangle$$

$$\implies \ E[\phi(x)] = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) =: \mu$$

Solving the maximum likelihood problem is **easy**.
Estimate the decay constant of an atom:
We use exponential family notation where

\[ p(x; \theta) = \exp(\langle -x, \theta \rangle - (- \log \theta)) \]

**Computing \( \mu \)**
Since \( \phi(x) = -x \) all we need to do is **average over all decay times** that we observe.

**Solving for Maximum Likelihood**
The maximum likelihood condition yields

\[ \mu = \partial_\theta g(\theta) = \partial_\theta(- \log \theta) = - \frac{1}{\theta} \]

This leads to \( \theta = - \frac{1}{\mu} \).
Entropy

Basically it’s the number of bits needed to encode a random variable. It is defined as

$$H(p) = \int -p(x) \log p(x) \, dx$$

where we set $0 \log 0 := 0$

Maximum Entropy Density

The density $p(x)$ satisfying $\mathbb{E}[\phi(x)] \geq \eta$ with maximum entropy is $\exp(\langle \phi(x), \theta \rangle - g(\theta))$.

Corollary

The most vague density with a given variance is the Gaussian distribution.

Corollary

The most vague density with a given mean is the Laplacian distribution.
Observe Data
\[ x_1, \ldots, x_m \text{ drawn from distribution } p(x | \theta) \]

Compute Likelihood
\[
p(X | \theta) = \prod_{i=1}^{m} \exp(\langle \phi(x_i), \theta \rangle - g(\theta))
\]

Maximize it
Take the negative log and minimize, which leads to
\[
\partial_\theta g(\theta) = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i)
\]

This can be solved analytically or (whenever this is impossible or we are lazy) by Newton’s method.
Simple Data

Discrete random variables (e.g. tossing a dice).

<table>
<thead>
<tr>
<th>Outcome</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counts</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Probabilities</td>
<td>0.15</td>
<td>0.30</td>
<td>0.10</td>
<td>0.05</td>
<td>0.20</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Maximum Likelihood Solution

Count the number of outcomes and use the relative frequency of occurrence as estimates for the probability:

\[
p_{\text{emp}}(x) = \frac{\#x}{m}
\]

Problems

- Bad idea if we have few data.
- Bad idea if we have continuous random variables.
Tossing a dice
Fisher Information and Efficiency

Fisher Score

\[ V_\theta(x) := \partial_\theta \log p(x; \theta) \]

This tells us the influence of \( x \) on estimating \( \theta \). Its expected value vanishes, since

\[
E[\partial_\theta \log p(X; \theta)] = \int p(X; \theta) \partial_\theta \log p(X; \theta) dX
\]

\[ = \partial_\theta \int p(X; \theta) dX = 0. \]

Fisher Information Matrix

It is the covariance matrix of the Fisher scores, that is

\[ I := \text{Cov}[V_\theta(x)] \]
Cramer Rao Theorem

**Efficiency**

Covariance of estimator $\hat{\theta}(X)$ rescaled by $I$:

$$1/e := \det \text{Cov}[\hat{\theta}(X)]\text{Cov}[\partial_{\theta} \log p(X; \theta)]$$

**Theorem**

The efficiency for unbiased estimators is never better (i.e. larger) than 1. Equality is achieved for MLEs.

**Proof (scalar case only)**

By Cauchy-Schwartz we have

$$\left( \mathbb{E}_\theta \left[ (V_{\theta}(X) - \mathbb{E}_\theta [V_{\theta}(X)]) \left( \hat{\theta}(X) - \mathbb{E}_\theta \left[ \hat{\theta}(X) \right] \right) \right] \right)^2$$

$$\leq \mathbb{E}_\theta \left[ (V_{\theta}(X) - \mathbb{E}_\theta [V_{\theta}(X)])^2 \right] \mathbb{E}_\theta \left[ \left( \hat{\theta}(X) - \mathbb{E}_\theta \left[ \hat{\theta}(X) \right] \right)^2 \right] = IB$$
Proof

At the same time, \( \mathbb{E}_\theta [V_\theta(X)] = 0 \) implies that

\[
\mathbb{E}_\theta \left[ (V_\theta(X) - \mathbb{E}_\theta [V_\theta(X)]) \left( \hat{\theta}(X) - \mathbb{E}_\theta \left[ \hat{\theta}(X) \right] \right) \right]
\]

\[
= \mathbb{E}_\theta \left[ V_\theta(X) \hat{\theta}(X) \right]
\]

\[
= \left( \int p(X|\theta) \partial_\theta \log p(X|\theta) \hat{\theta}(X) dX \right)
\]

\[
= \partial_\theta \int p(X|\theta) \hat{\theta}(X) dX = \partial_\theta \theta = 1.
\]

Cautionary Note

This does not imply that a biased estimator might not have lower variance.
Fisher Score

\[ V_\theta(x) = \partial_\theta \log p(x; \theta) \]
\[ = \phi(x) - \partial_\theta g(\theta) \]

Fisher Information

\[ I = \text{Cov}[V_\theta(x)] \]
\[ = \text{Cov}[\phi(x) - \partial_\theta g(\theta)] \]
\[ = \partial_{\theta}^2 g(\theta) \]

Efficiency of estimator can be obtained directly from log-partition function.

Outer Product Matrix

It is given (up to an offset) by \( \langle \phi(x), \phi(x') \rangle \). This leads to Kernel-PCA . . .
Problems with Maximum Likelihood
With not enough data, parameter estimates will be bad.

Prior to the rescue
Often we know where the solution should be. So we encode the latter by means of a prior $p(\theta)$.

Normal Prior
Simply set $p(\theta) \propto \exp\left(-\frac{1}{2\sigma^2}\|\theta\|^2\right)$.

Posterior

$$p(\theta|X) \propto \exp\left(\sum_{i=1}^{m}\langle \phi(x_i), \theta \rangle - g(\theta) - \frac{1}{2\sigma^2}\|\theta\|^2\right)$$
Tossing a dice with priors
Conjugate Priors

Problem with Normal Prior
The posterior looks different from the likelihood. So many of the Maximum Likelihood optimization algorithms may not work ...

Idea
What if we had a prior which looked like additional data, that is

\[ p(\theta | X) \sim p(X | \theta) \]

For exponential families this is easy. Simply set

\[ p(\theta | a) \propto \exp(\langle \theta, m_0 a \rangle - m_0 g(\theta)) \]

Posterior

\[ p(\theta | X) \propto \exp \left( (m + m_0) \left( \left\langle \frac{m \mu + m_0 a}{m + m_0}, \theta \right\rangle - g(\theta) \right) \right) \]
Laplace Rule
A conjugate prior with parameters \((a, m_0)\) in the multinomial family could be to set \(a = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)\). This is often also called the **Dirichlet prior**. It leads to

\[
p(x) = \frac{\#x + m_0/n}{m + m_0} \text{ instead of } p(x) = \frac{\#x}{m}
\]

Example

<table>
<thead>
<tr>
<th>Outcome</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counts</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>MLE</td>
<td>0.15</td>
<td>0.30</td>
<td>0.10</td>
<td>0.05</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>MAP ((m_0 = 6))</td>
<td>0.25</td>
<td>0.27</td>
<td>0.12</td>
<td>0.08</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>MAP ((m_0 = 100))</td>
<td>0.16</td>
<td>0.19</td>
<td>0.16</td>
<td>0.15</td>
<td>0.17</td>
<td>0.17</td>
</tr>
</tbody>
</table>
Optimization Problems

Maximum Likelihood

\[
\minimize_{\theta} \sum_{i=1}^{m} g(\theta) - \langle \phi(x_i), \theta \rangle \quad \Longrightarrow \quad \partial_\theta g(\theta) = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i)
\]

Normal Prior

\[
\minimize_{\theta} \sum_{i=1}^{m} g(\theta) - \langle \phi(x_i), \theta \rangle + \frac{1}{2\sigma^2} \|\theta\|^2
\]

Conjugate Prior

\[
\minimize_{\theta} \sum_{i=1}^{m} g(\theta) - \langle \phi(x_i), \theta \rangle + m_0 g(\theta) - m_0 \langle \tilde{\mu}, \theta \rangle
\]

equivalently solve \[
\partial_\theta g(\theta) = \frac{1}{m + m_0} \sum_{i=1}^{m} \phi(x_i) + \frac{m_0}{m + m_0} \tilde{\mu}
\]
Summary

Model
- Log partition function
- Expectations and derivatives
- Maximum entropy formulation

A Zoo of Densities

Estimation
- Maximum Likelihood Estimator
- Fisher Information Matrix and Cramer Rao Theorem
- Normal Priors and Conjugate Priors
- Fisher information and log-partition function
Exponential Families and Kernels
Lecture 2

Alexander J. Smola
Alex.Smola@nicta.com.au

Machine Learning Program
National ICT Australia
RSISE, The Australian National University
Outline

Exponential Families
- Maximum likelihood and Fisher information
- Priors (conjugate and normal)

Conditioning and Feature Spaces
- Conditional distributions and inner products
- Clifford Hammersley Decomposition

Applications
- Classification and novelty detection
- Regression

Applications
- Conditional random fields
- Intractable models and semidefinite approximations
Clifford Hammersley Theorem and Graphical Models
- Decomposition results
- Key connection

Conditional Distributions
- Log partition function
- Expectations and derivatives
- Inner product formulation and kernels
- Gaussian Processes

Applications
- Classification + Regression
- Conditional Random Fields
- Spatial Poisson Models
**Conditional Independence**

- $x, x'$ are conditionally independent given $c$, if

$$p(x, x'|c) = p(x|c)p(x'|c)$$

- Distributions can be simplified greatly by conditional independence assumptions.

**Markov Network**

- Given a graph $G(V, E)$ with vertices $V$ and edges $E$ associate a random variable $x \in \mathbb{R}^{|V|}$ with $G$.

- Subsets of random variables $x_S, x_{S'}$ are conditionally independent given $x_C$ if removing the vertices $C$ from $G(V, E)$ decomposes the graph into disjoint subsets containing $S, S'$. 
Conditional Independence
Clques

Definition

- Subset of the graph which is fully connected
- Maximal Cliques (they define the graph)

Advantage

- Easy to specify dependencies between variables
- Use graph algorithms for inference
Hammersley Clifford Theorem

**Problem**
Specify $p(x)$ with conditional independence properties.

**Theorem**

$$p(x) = \frac{1}{Z} \exp \left( \sum_{c \in C} \psi_c(x_c) \right)$$

whenever $p(x)$ is nonzero on the entire domain.

**Application**
Apply decomposition for exponential families where

$$p(x) = \exp(\langle \phi(x), \theta \rangle - g(\theta)).$$

**Corollary**
The sufficient statistics $\phi(x)$ decompose according to

$$\phi(x) = (\ldots, \phi_c(x_c), \ldots) \implies \langle \phi(x), \phi(x') \rangle = \sum_c \langle \phi_c(x_c), \phi_c(x'_c) \rangle$$
Proof

**Step 1: Obtain linear functional**

Combing the exponential setting with the CH theorem:

\[
\langle \Phi(x), \theta \rangle = \sum_{c \in C} \psi_c(x_c) - \log Z + g(\theta) \text{ for all } x, \theta.
\]

**Step 2: Orthonormal basis in } \theta**

Pick an orthonormal basis and swallow \( Z, g \). This gives

\[
\langle \Phi(x), e_i \rangle = \sum_{c \in C} \eta^i_c(x_c) \text{ for some } \eta^i_c(x_c).
\]

**Step 3: Reconstruct sufficient statistics**

\[
\Phi_c(x_c) := (\eta^1_c(x_c), \eta^2_c(x_c), \ldots)
\]

which allows us to compute

\[
\langle \Phi(x), \theta \rangle = \sum_{c \in C} \sum_i \theta_i \Phi^i_c(x_c).
\]
Sufficient Statistics
Recall that for normal distributions $\phi(x) = (x, xx^\top)$.

Clifford Hammersley Application
- $\phi(x)$ must decompose into subsets involving only variables from each maximal clique.
- The linear term $x$ is OK by default.
- The only nonzero terms coupling $x_i x_j$ are those corresponding to an edge in the graph $G(V, E)$.

Inverse Covariance Matrix
- The natural parameter aligned with $xx^\top$ is the inverse covariance matrix.
- Its sparsity mirrors $G(V, E)$.
- Hence a sparse inverse kernel matrix corresponds to graphical model!
Density

\[
p(x|\theta) = \exp \left( \sum_{i=1}^{n} x_i \theta_{1i} + \sum_{i,j=1}^{n} x_i x_j \theta_{2ij} - g(\theta) \right)
\]

Here \( \theta_2 = \Sigma^{-1} \), is the inverse covariance matrix. We have that \( (\Sigma^{-1})_{ij} \neq 0 \) only if \((i, j)\) share an edge.
Conditional Distributions

Conditional Density

\[ p(x|\theta) = \exp(\langle \phi(x), \theta \rangle - g(\theta)) \]
\[ p(y|x, \theta) = \exp(\langle \phi(x, y), \theta \rangle - g(\theta|x)) \]

Log-partition function

\[ g(\theta|x) = \log \int_y \exp(\langle \phi(x, y), \theta \rangle) dy \]

Sufficient Criterion

\[ p(x, y|\theta) \] is a member of the exponential family itself.

Key Idea

Avoid computing \( \phi(x, y) \) directly, only evaluate inner products via

\[ k((x, y), (x', y')) := \langle \phi(x, y), \phi(x', y') \rangle \]
Conditional Distributions

Maximum a Posteriori Estimation

\[-\log p(\theta|X) = \sum_{i=1}^{m} -\langle \phi(x_i), \theta \rangle + mg(\theta) + \frac{1}{2\sigma^2} \|\theta\|^2 + c\]

\[-\log p(\theta|X, Y) = \sum_{i=1}^{m} -\langle \phi(x_i, y_i), \theta \rangle + g(\theta|x_i) + \frac{1}{2\sigma^2} \|\theta\|^2 + c\]

Solving the Problem

- The problem is strictly convex in $\theta$.
- Direct solution is impossible if we cannot compute $\phi(x, y)$ directly.
- Solve convex problem in expansion coefficients.
- Expand $\theta$ in a linear combination of $\phi(x_i, y_i)$.
Joint Feature Map

\[ \phi(x, y) \]

Evaluate for all \( y \)
Representer Theorem

Objective Function

\[- \log p(\theta | X, Y) = \sum_{i=1}^{m} -\langle \phi(x_i, y_i), \theta \rangle + g(\theta | x_i) + \frac{1}{2\sigma^2} \| \theta \|^2 + c \]

Decomposition

- Decompose \( \theta \) into \( \theta = \theta_{\parallel} + \theta_{\perp} \) where
  \( \theta_{\parallel} \in \text{span}\{ \phi(x_i, y) \text{ where } 1 \leq i \leq m \text{ and } y \in Y \} \)

- Both \( g(\theta | x_i) \) and \( \langle \phi(x_i, y_i), \theta \rangle \) are independent of \( \theta_{\perp} \).

Theorem

\(- \log p(\theta | X, Y) \) is minimized for \( \theta_{\perp} = 0 \), hence \( \theta = \theta_{\parallel} \).

Consequence

If \( \text{span}\{ \phi(x_i, y) \text{ where } 1 \leq i \leq m \text{ and } y \in Y \} \) is finite dimensional, we have a parametric optimization problem.
Expansion

\[ \theta = \sum_{i=1}^{m} \sum_{y \in Y} \alpha_{iy} \phi(x_i, y) \]

Inner Product

\[ \langle \phi(x, y), \theta \rangle = \sum_{i=1}^{m} \sum_{y \in Y} \alpha_{iy} k((x, y), (x_i, y)) \]

Norm

\[ \| \theta \|^2 = \sum_{i,j=1}^{m} \sum_{y, y' \in Y} \alpha_{iy} \alpha_{jy'} k((x_i, y), (x_j, y')) \]

Log-partition function

\[ g(\theta | x) = \log \sum_{y \in Y} \exp (\langle \phi(x, y), \theta \rangle) \]
Normal Prior on $\theta$ . . .

$\theta \sim \mathcal{N}(0, \sigma^2 1)$

. . . yields Normal Prior on $t(x, y) = \langle \phi(x, y), \theta \rangle$

- Distribution of projected Gaussian is Gaussian.
- The mean vanishes

$$E_{\theta}[t(x, y)] = \langle \phi(x, y), E_{\theta}[\theta] \rangle = 0$$

- The covariance yields

$$\text{Cov}[t(x, y), t(x', y')] = E_{\theta} [\langle \phi(x, y), \theta \rangle \langle \theta, \phi(x', y') \rangle]$$

$$= \sigma^2 \langle \phi(x, y), \phi(x', y') \rangle$$

$$:= k((x, y), (x', y'))$$

. . . so we have a Gaussian Process on $x$ . . .

with kernel $k((x, y), (x', y')) = \sigma^2 \langle \phi(x, y), \phi(x', y') \rangle$. 
Linear Covariance

\begin{center}
\begin{tikzpicture}
\begin{axis}[
    xlabel={y},
    ylabel={$k(x,y)$ for $x=1$},
    xmin=-5, xmax=5,
    ymin=-5, ymax=5,
    grid=both,
    axis lines=middle,
]
\addplot[domain=-5:5,samples=100] {x};
\end{axis}
\end{tikzpicture}
\end{center}
Laplacian Covariance

\[ k(x,y) \text{ for } x=1 \]

The graph shows the Laplacian covariance function for a fixed value of \( x = 1 \). The function is symmetric around \( y = 0 \) and exhibits a sharp peak at \( y = 0 \), with values decreasing as \( y \) moves away from 0.
Gaussian Covariance
Polynomial (Order 3)
$B_3$-Spline Covariance
Sample from Gaussian RBF
Sample from Gaussian RBF
Sample from Gaussian RBF
Sample from Gaussian RBF
Sample from linear kernel
Sample from linear kernel
Sample from linear kernel
Sample from linear kernel
Sample from linear kernel
Choose a suitable sufficient statistic $\phi(x, y)$

Conditionally multinomial distribution leads to Gaussian Process multiclass estimator: we have a distribution over $n$ classes which depends on $x$.

Conditionally Gaussian leads to Gaussian Process regression: we have a normal distribution over a random variable which depends on the location.

**Note:** we estimate mean and variance.

Conditionally Poisson distributions yield locally varying Poisson processes. This has no name yet ...

Solve the optimization problem

This is typically convex.

The bottom line

Instead of choosing $k(x, x')$ choose $k((x, y), (x', y'))$. 
Example: GP Classification

**Sufficient Statistic**

We pick $\phi(x, y) = \phi(x) \otimes e_y$, that is

$$k((x, y), (x', y')) = k(x, x') \delta_{yy'} \text{ where } y, y' \in \{1, \ldots, n\}$$

**Kernel Expansion**

By the representer theorem we get that

$$\theta = \sum_{i=1}^{m} \sum_{y} \alpha_{iy} \phi(x_i, y)$$

**Optimization Problem**

Big mess . . . but convex.
A Toy Example
Noisy Data
Summary

Clifford Hammersley Theorem and Graphical Models
- Decomposition results
- Key connection
- Normal distribution

Conditional Distributions
- Log partition function
- Expectations and derivatives
- Inner product formulation and kernels
- Gaussian Processes

Applications
- Generalized kernel trick
- Conditioning gives existing estimation methods back
Outline

Exponential Families
- Maximum likelihood and Fisher information
- Priors (conjugate and normal)

Conditioning and Feature Spaces
- Conditional distributions and inner products
- Clifford Hammersley Decomposition

Applications
- Classification and novelty detection
- Regression

Applications
- Conditional random fields
- Intractable models and semidefinite approximations
Novelty Detection
- Density estimation
- Thresholding and likelihood ratio

Classification
- Log partition function
- Optimization problem
- Examples
- Clustering and transduction

Regression
- Conditional normal distribution
- Estimating the covariance
- Heteroscedastic estimators
Density Estimation

**Maximum a Posteriori**

\[
\minimize_{\theta} \sum_{i=1}^{m} g(\theta) - \langle \phi(x_i), \theta \rangle + \frac{1}{2\sigma^2} \|\theta\|^2
\]

**Advantages**
- Convex optimization problem
- Concentration of measure

**Problems**
- Normalization \( g(\theta) \) may be painful to compute
- For density estimation we need no normalized \( p(x|\theta) \)
- No need to perform particularly well in high density regions
Novelty Detection

Optimization Problem

**MAP**
\[
\sum_{i=1}^{m} - \log p(x_i | \theta) + \frac{1}{2\sigma^2} \|\theta\|^2
\]

**Novelty**
\[
\sum_{i=1}^{m} \max \left( - \log \frac{p(x_i | \theta)}{\exp(\rho - g(\theta))}, 0 \right) + \frac{1}{2} \|\theta\|^2
\]
\[
\sum_{i=1}^{m} \max(\rho - \langle \phi(x_i), \theta \rangle, 0) + \frac{1}{2} \|\theta\|^2
\]

Advantages

- No normalization \( g(\theta) \) needed
- No need to perform particularly well in high density regions (estimator focuses on low-density regions)
- Quadratic program
Geometric Interpretation

Idea
Find hyperplane that has maximum distance from origin, yet is still closer to the origin than the observations.

Hard Margin

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \| \theta \|^2 \\
\text{subject to} & \quad \langle \theta, x_i \rangle \geq 1
\end{align*}
\]

Soft Margin

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \| \theta \|^2 + C \sum_{i=1}^{m} \xi_i \\
\text{subject to} & \quad \langle \theta, x_i \rangle \geq 1 - \xi_i \\
& \quad \xi_i \geq 0
\end{align*}
\]
Primal Problem

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|\theta\|^2 + C \sum_{i=1}^{m} \xi_i \\
\text{subject to} & \quad \langle \theta, x_i \rangle - 1 + \xi_i \geq 0 \text{ and } \xi_i \geq 0
\end{align*}
\]

Lagrange Function

We construct a \textbf{Lagrange Function} \( L \) by subtracting the constraints, multiplied by \textbf{Lagrange multipliers} \( (\alpha_i \text{ and } \eta_i) \), from the \textbf{Primal Objective Function}. \( L \) has a \textbf{saddlepoint} at the optimal solution.

\[
L = \frac{1}{2} \|\theta\|^2 + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i (\langle \theta, x_i \rangle - 1 + \xi_i) - \sum_{i=1}^{m} \eta_i \xi_i
\]

where \( \alpha_i, \eta_i \geq 0 \). For instance, if \( \xi_i < 0 \) we could increase \( L \) without bound via \( \eta_i \).
Optimality Conditions

\[ \partial_{\theta} L = \theta - \sum_{i=1}^{m} \alpha_i x_i = 0 \implies \theta = \sum_{i=1}^{m} \alpha_i x_i \]

\[ \partial_{\xi_i} L = C - \alpha_i - \eta_i = 0 \implies \alpha_i \in [0, C] \]

Now we substitute the two optimality conditions back into \( L \) and eliminate the primal variables.

Dual Problem

minimize \( \frac{1}{2} \sum_{i=1}^{m} \alpha_i \alpha_j \langle x_i, x_j \rangle - \sum_{i=1}^{m} \alpha_i \)

subject to \( \alpha_i \in [0, C] \)

Convexity ensures uniqueness of the optimum.
The $\nu$-Trick

**Problem**

Depending on how we choose $C$, the number of points selected as lying on the “wrong” side of the hyperplane $H := \{x | \langle \theta, x \rangle = 1 \}$ will vary.

- We would like to specify a certain fraction $\nu$ beforehand.
- We want to make the setting more adaptive to the data.

**Solution**

Use adaptive hyperplane that separates data from the origin, i.e. find

$$H := \{x | \langle \theta, x \rangle = \rho \},$$

where the threshold $\rho$ is adaptive.
Primal Problem

minimize \[ \frac{1}{2} \| \theta \|^2 + \sum_{i=1}^{m} \xi_i - m \nu \rho \]
subject to \[ \langle \theta, x_i \rangle - \rho + \xi_i \geq 0 \text{ and } \xi_i \geq 0 \]

Dual Problem

minimize \[ \frac{1}{2} \sum_{i=1}^{m} \alpha_i \alpha_j \langle x_i, x_j \rangle \]
subject to \[ \alpha_i \in [0, 1] \text{ and } \sum_{i=1}^{m} \alpha_i = \nu m. \]

Difference to before
The \( \sum_i \alpha_i \) term vanishes from the objective function but we get one more constraint, namely \( \sum_i \alpha_i = \nu m. \)
The $\nu$-Property

Optimization Problem

\[
\text{minimize} \quad \frac{1}{2} ||\theta||^2 + \sum_{i=1}^{m} \xi_i - m\nu\rho
\]

subject to \quad \langle \theta, x_i \rangle - \rho + \xi_i \geq 0 \quad \text{and} \quad \xi_i \geq 0

Theorem

- At most a fraction of $\nu$ points will lie on the “wrong” side of the margin, i.e., $y_i f(x_i) < 1$.
- At most a fraction of $1 - \nu$ points will lie on the “right” side of the margin, i.e., $y_if(x_i) > 1$.
- In the limit, those fractions will become exact.

Proof Idea

At optimum, shift $\rho$ slightly: only the active constraints will have an influence on the objective function.
## Classification

### Maximum a Posteriori Estimation

\[
- \log p(\theta | X, Y) = \sum_{i=1}^{m} -\langle \phi(x_i, y_i), \theta \rangle + g(\theta | x_i) + \frac{1}{2\sigma^2} \| \theta \|^2 + c
\]

### Domain
- Finite set of observations \( Y = \{1, \ldots, m\} \)
- Log-partition function \( g(\theta | x) \) easy to compute.
- Optional centering

\[
\phi(x, y) \rightarrow \phi(x, y) + c
\]

leaves \( p(y | x, \theta) \) unchanged (offsets both terms).

### Gaussian Process Connection
- Inner product \( t(x, y) = \langle \phi(x, y), \theta \rangle \) is drawn from Gaussian process, so same setting as in literature.
Classification

**Sufficient Statistic**

We pick $\phi(x, y) = \phi(x) \otimes e_y$, that is

$$k((x, y), (x', y')) = k(x, x')\delta_{yy'} \text{ where } y, y' \in \{1, \ldots, n\}$$

**Kernel Expansion**

By the representer theorem we get that

$$\theta = \sum_{i=1}^{m} \sum_y \alpha_{iy} \phi(x_i, y)$$

**Optimization Problem**

- Big mess . . . but convex.
- Solve by Newton or Block-Jacobi method.
Problems with GP Classification

- Optimize even where classification is good
- Only sign of classification needed
- Only "strongest" wrong class matters
- Want to classify with a margin

Optimization Problem

**MAP**
$$\sum_{i=1}^{m} - \log p(y_i | x_i, \theta) + \frac{1}{2\sigma^2} \|\theta\|^2$$

**SVM**
$$\sum_{i=1}^{m} \max\left(\rho - \log \frac{p(y_i | x_i, \theta)}{\max_{y \neq y_i} p(y | x_i, \theta)}, 0\right) + \frac{1}{2} \|\theta\|^2$$

$$\sum_{i=1}^{m} \max(\rho - \langle \phi(x_i, y_i), \theta \rangle + \max_{y \neq y_i} \langle \phi(x_i, y), \theta \rangle, 0) + \frac{1}{2} \|\theta\|^2$$
Sufficient Statistics

- Offset in $\phi(x, y)$ can be arbitrary
- Pick such that $\phi(x, y) = y\phi(x)$ where $y \in \{\pm 1\}$.
- Kernel matrix becomes

$$K_{ij} = k((x_i, y_i), (x_j, y_j)) = y_i y_j k(x_i, x_j)$$

Optimization Problem

- The max over other classes becomes

$$\max_{y \neq y_i} \langle \phi(x_i, y), \theta \rangle = -y \langle \phi(x_i), \theta \rangle$$

- Overall problem

$$\sum_{i=1}^{m} \max(\rho - 2y_i \langle \phi(x_i), \theta \rangle, 0) + \frac{1}{2} ||\theta||^2$$
Minimize \( \frac{1}{2} \| \theta \|^2 \) subject to \( y_i(\langle \theta, x_i \rangle + b) \geq 1 \) for all \( i \).
Optimization Problem

**Linear Function**

\[ f(x) = \langle \theta, x \rangle + b \]

**Mathematical Programming Setting**

If we require error-free classification with a margin, i.e., \( y f(x) \geq 1 \), we obtain:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \| \theta \|^2 \\
\text{subject to} & \quad y_i (\langle \theta, x_i \rangle + b) - 1 \geq 0 \text{ for all } 1 \leq i \leq m
\end{align*}
\]

**Result**

The dual of the optimization problem is a simple quadratic program (more later ...).

**Connection back to conditional probabilities**

Offset \( b \) takes care of bias towards one of the classes.
Regression

Maximum a Posteriori Estimation

\[- \log p(\theta|X, Y) = \sum_{i=1}^{m} -\langle \phi(x_i, y_i), \theta \rangle + g(\theta|x_i) + \frac{1}{2\sigma^2} \|\theta\|^2 + c \]

Domain

- Continuous domain of observations $y = \mathbb{R}$
- Log-partition function $g(\theta|x)$ easy to compute in closed form as normal distribution.

Gaussian Process Connection

Inner product $t(x, y) = \langle \phi(x, y), \theta \rangle$ is drawn from Gaussian process. In particular also rescaled mean and covariance.
Regression

Sufficient Statistic (Standard Model)
We pick $\phi(x, y) = (y\phi(x), y^2)$, that is

$$k((x, y), (x', y')) = k(x, x')yy' + y^2y'^2 \text{ where } y, y' \in \mathbb{R}$$

Traditionally the variance is fixed, that is $\theta_2 = \text{const.}$.

Sufficient Statistic (Fancy Model)
We pick $\phi(x, y) = (y\phi_1(x), y^2\phi_2(x))$, that is

$$k((x, y), (x', y')) = k_1(x, x')yy' + k_2(x, x')y^2y'^2 \text{ where } y, y' \in \mathbb{R}$$

We estimate mean and variance simultaneously.

Kernel Expansion
By the representer theorem (and more algebra) we get

$$\theta = \left( \sum_{i=1}^{m} \alpha_i \phi_1(x_i), \sum_{i=1}^{m} \alpha_i \phi_2(x_i) \right)$$
Training Data
\[ \langle k^\top (x)(K + \sigma^2 1)^{-1}y \]

Variance \( k(x, x) + \sigma^2 - \overrightarrow{k}^\top(x)(K + \sigma^2 1)^{-1}\overrightarrow{k}(x) \)
Optimization Problem:

\[
\begin{align*}
\minimize \sum_{i=1}^{m} & \left[ -\frac{1}{4} \left( \sum_{j=1}^{m} \alpha_{1j} k_{1}(x_i, x_j) \right) \right] \top \left[ \sum_{j=1}^{m} \alpha_{2j} k_{2}(x_i, x_j) \right]^{-1} \left[ \sum_{j=1}^{m} \alpha_{1j} k_{1}(x_i, x_j) \right] \\
- \frac{1}{2} \log \det & -2 \left[ \sum_{j=1}^{m} \alpha_{2j} k_{2}(x_i, x_j) \right] - \sum_{j=1}^{m} \left[ y_i \top \alpha_{1j} k_{1}(x_i, x_j) + (y_j \top \alpha_{2j} y_j) k_{2}(x_i, x_j) \right] \\
+ \frac{1}{2\sigma^2} & \sum_{i,j} \alpha_{1i} \alpha_{1j} k_{1}(x_i, x_j) + \text{tr} \left[ \alpha_{2i} \alpha_{2j} \right] k_{2}(x_i, x_j).
\end{align*}
\]

subject to \( 0 \succ \sum_{i=1}^{m} \alpha_{2i} k(x_i, x_j) \)

Properties of the problem:

- The problem is convex
- The log-determinant from the normalization of the Gaussian acts as a barrier function.
- We get a semidefinite program.
Heteroscedastic Regression

regression estimation and training data

variance estimation
θ1 estimation

θ2 estimation
Novelty Detection
- Density estimation
- Thresholding and likelihood ratio

Classification
- Log partition function
- Optimization problem
- Examples
- Clustering and transduction

Regression
- Conditional normal distribution
- Estimating the covariance
- Heteroscedastic estimators
Exponential Families and Kernels

Lecture 4

Alexander J. Smola
Alex.Smola@nicta.com.au

Machine Learning Program
National ICT Australia
RSISE, The Australian National University
Outline

Exponential Families
- Maximum likelihood and Fisher information
- Priors (conjugate and normal)

Conditioning and Feature Spaces
- Conditional distributions and inner products
- Clifford Hammersley Decomposition

Applications
- Classification and novelty detection
- Regression

Applications
- Conditional random fields
- Intractable models and semidefinite approximations
Conditional Random Fields
- Structured random variables
- Subspace representer theorem and decomposition
- Derivatives and conditional expectations

Inference and Message Passing
- Dynamic programming
- Message passing and junction trees
- Intractable cases

Semidefinite Relaxations
- Marginal polytopes
- Fenchel duality and entropy
- Relaxations for conditional random fields
Decomposition

The sufficient statistics $\phi(x)$ decompose according to

$$\phi(x) = (\ldots, \phi_c(x_c), \ldots)$$

Consequently we can write the kernel via

$$k(x, x') = \langle \phi(x), \phi(x') \rangle = \sum_c \langle \phi_c(x_c), \phi_c(x'_c) \rangle = \sum_c k_c(x_c, x'_c)$$
Key Points

- Cliques are \((x_t, y_t), (x_t, x_{t+1}),\) and \((y_t, y_{t+1})\)
- We can drop cliques in \((x_t, x_{t+1})\): they do not affect

\[
p(y|x, \theta) = \exp \left( \sum_t \langle \phi_{xy}(x_t, y_t), \theta_{xy,t} \rangle + \langle \phi_{yy}(y_t, y_{t+1}), \theta_{yy,t} \rangle + \langle \phi_{xx}(x_t, x_{t+1}), \theta_{xx,t} \rangle - g(\theta|x) \right)
\]
Computational Issues

Key Points

- Compute $g(\theta|x)$ via dynamic
- Assume stationarity of the model, that is $\theta_c$ does not depend on the position of the

Dynamic Programming

$$g(\theta|x) = \log \sum_{y_1,\ldots,y_T} \prod_{t=1}^{T} \exp \left( \langle \phi_{xy}(x_t, y_t), \theta_{xy} \rangle + \langle \phi_{yy}(y_t, y_{t+1}), \theta_{yy} \rangle \right)$$

$$= \log \sum_{y_1} \sum_{y_2} M_1(y_1, y_2) \sum_{y_3} M_2(y_2, y_3) \ldots \sum_{y_T} M_T(y_{T-1}, y_T)$$

So we can compute $g(\theta|x), p(y_t|x, \theta)$ and $p(y_t, y_{t+1}|x, \theta)$ via dynamic programming.
Forward Backward Algorithm

Key Idea

- Store sum over all $y_1, \ldots, y_{t-1}$ (forward pass) and over all $y_{t+1}, \ldots, y_T$ as intermediate values.
- We get those values for all positions $t$ in one sweep.
- Extend this to message passing (when we have trees).
Minimization

Objective Function

\[- \log p(\theta | X, Y) = \sum_{i=1}^{m} - \langle \phi(x_i, y_i), \theta \rangle + g(\theta | x_i) + \frac{1}{2\sigma^2} \| \theta \|^2 + c\]

\[\partial_{\theta} - \log p(\theta | X, Y) = \sum_{i=1}^{m} - \phi(x_i, y_i) + E [\phi(x_i, y_i) | x_i] + \frac{1}{\sigma^2} \theta\]

We only need \(E [\phi_{xy}(x_{it}, y_{it}) | x_i]\) and \(E [\phi_{yy}(y_{it}, y_{i(t+1)}) | x_i]\).

Kernel Trick

![bullet] Conditional expectations of \(\Phi(x_{it}, y_{it})\) cannot be computed explicitly but inner products can.

\[\langle \phi_{xy}(x'_t, y'_t), E [\phi_{xy}(x_t, y_t) | x] = E [k((x'_t, y'_t), (x_t, y_t) | x]\]

![bullet] Only need marginals \(p(y_t | x, \theta)\) and \(p(y_t, y_{t+1} | x, \theta)\), which we get via dynamic programming.
Representer Theorem

Solutions of the MAP problem are given by

$$\theta \in \text{span}\{\phi(x_i, y) \text{ for all } y \in \mathcal{Y} \text{ and } 1 \leq i \leq n\}$$

Big Problem

$|\mathcal{Y}|$ could be huge, e.g. for sequence annotation $2^n$.

Solution

- Exploit decomposition of $\phi(x, y)$ into sufficient statistics on cliques.
- Restriction of $\mathcal{Y}$ to cliques is much smaller.

$$\theta_c \in \text{span}\{\phi_c(x_{ci}, y_c) \text{ for all } y_c \in \mathcal{Y}_c \text{ and } 1 \leq i \leq n\}$$

Rather than $2^n$ we now get $2^{|c|}$. 
Conditional Random Field: maximize $p(y|x, \theta)$

Hidden Markov Model: maximize $p(x, y|\theta)$
Equivalence Theorem

**Theorem**

CRFs and HMMs yield identical probability estimates for $p(y|x, \theta)$, if the set of functions is equally expressive.

**Proof**

- Write out $p_{\text{CRF}}(y|x, \theta)$ and $p_{\text{HMM}}(x, y|\theta)$, and show that they only differ in the normalization.
- This disappears when computing $p_{\text{HMM}}(y|x, \theta)$.

**Consequence**

Differential training for current HMM implementations.
Message Passing
Message Passing

Idea
Extend the forward-backward idea to trees.

Algorithm
- Given clique potentials $M(y_i, y_j)$
- Initialize messages $\mu_{ij}(y_j) = 1$
- Update outgoing messages by

$$\mu_{ij}(y_j) = \sum_{y_i \in Y_i} \prod_{k \neq j} \mu_{ki}(y_i) M_{ij}(y_i, y_j)$$

Here $(i, k)$ is an edge in the graph.

Theorem
The message passing algorithm converges after $n$ iterations ($n$ is diameter of graph).

Hack
Use this for graphs with loops and hope . . .
Stock standard algorithms available to transform graph into junction tree. Now we can use message passing . . .
Junction Tree Algorithm

Idea
Messages involve variables in the separator sets.

Algorithm
- Given clique potentials $M_c(y_c)$ and separator sets $s$.
- Initialize messages $\mu_{c,s}(y_s) = 1$.
- Update outgoing messages by

$$
\mu_{c,s}(y_s) = \sum_{y_c \setminus y_s} \prod_{s' \neq s} \mu_{c',s'}(y_{s'}) M_c(y_c)
$$

Here $s'$ is a separator set connecting $c$ with $c'$.

Theorem
The message passing algorithm converges after $n$ iterations ($n$ is diameter of the hypergraph).

Hack
Use this for graphs with loops and hope . . .
Scaling
The algorithm scales exponentially in the treewidth. Messages are of size $d|y_s|$.

Convergence with loops
Use of message passing may or may not converge. No real proof available.

Workaround
Use a subset of the graph and solve the inference problem with this. Average over spanning trees.

Workaround
Use sampling methods for inference.
**Fenchel Duality**

Compute dual of log-partition function via

\[ g^*(\mu) = \sup_{\theta \in \Theta} \langle \mu, \theta \rangle - g(\theta) \quad (\Theta \text{ is a convex domain}) \]

**Entropy and Expectation Parameters**

The maximum of the optimization problem is obtained for \( \mu = \partial_\theta g(\theta) \). This leads to

\[ H = \int -\log p(x|\theta)p(x|\theta) d\theta = -\langle \mu(\theta), \theta \rangle + g(\theta) = -g^*(\mu) \]

**Strong Duality**

Dualizing again leads to

\[ g(\theta) = \sup_{\mu \in M} \langle \theta, \mu \rangle + H(\mu) \]
Optimization Problem

\[ g(\theta) = \sup_{\mu \in M} \langle \theta, \mu \rangle + H(\mu) \]

Here \( M \) is the set of all possible marginals.

Relaxations on \( M \)

The polytope \( M \) is convex (by duality), however it is hard to compute (as hard as \( g(\theta) \)). So we relax it to \( \tilde{M} \) by impose constraints on higher order moments, such as

- Interval and linear inequality constraints.
- SDP constraints on the covariance matrix.

Upper bound on \( H(\mu) \)

Gaussian bound on the covariance via \( G(\mu) \). So we get

\[ g(\theta) \leq \sup_{\mu \in \tilde{M}} \langle \theta, \mu \rangle + G(\mu) \]
Optimization Problem

\[- \log p(\theta|X,Y) = \sum_{i=1}^{m} -\langle \phi(x_i, y_i), \theta \rangle + g(\theta|x_i) + \frac{1}{2\sigma^2} \|\theta\|^2 + c\]

\[\leq \sum_{i=1}^{m} \sup_{\mu_i \in \tilde{M}_i} \langle \theta, \mu_i - \phi(x_i, y_i) \rangle + H(\mu_i) + \frac{1}{2\sigma^2} \|\theta\|^2\]

Technical Details

- Minimization over \(\theta\) and \(\mu_i\) can be swapped (saddle-point property of a convex-concave problem) to obtain dual problem in \(\theta\).
- Map from \(\mu\) to moments in \(y|x\) via invertible sufficient statistics map.
- Constrained max-det problem.
Summary

Conditional Random Fields
- Structured random variables
- Subspace representer theorem and decomposition
- Derivatives and conditional expectations

Inference and Message Passing
- Dynamic programming
- Message passing and junction trees
- Intractable cases

Semidefinite Relaxations
- Marginal polytopes
- Fenchel duality and entropy
- Relaxations for conditional random fields
We are hiring. For details contact
Alex.Smola@nicta.com.au (http://www.nicta.com.au)

Positions
- PhD scholarships
- Postdoctoral positions, Senior researchers
- Long-term visitors (sabbaticals etc.)

More details on kernels
- http://www.kernel-machines.org
  Schölkopf and Smola: Learning with Kernels

Machine Learning Summer School
- http://www.mlss.cc
- MLSS’05 Canberra, Australia, 23/1-5/2/2005