Tractable Inference for Probabilistic Models

Manfred Opper

(Aston University, Birmingham, U.K.)

collaboration with:

Ole Winther (TU Denmark)

Dörthe Malzahn (TU Denmark)

Lehel Csató (Aston U)
The general Structure

\( D = \) Observed data

\( S = \) Hidden variables (unknown causes, etc)

Bayes Rule

\[
P(S|D) = \frac{P(D|S) \times P(S)}{P(D)}
\]

- \( P(S|D) \): posterior distribution
- \( P(D|S) \): Likelihood
- \( P(S) \): prior distribution
- \( P(D) \): Evidence or marginal likelihood
Overview

• Inference with probabilistic models: Examples

• A “canonical” model

• Problems with inference and approximate solutions

• Cavity/TAP approximation

• Applications

• Outlook
Example I: Modeling with Gaussian Processes

- **Observations**: Data $D = (y_1 \ldots, y_N)$ observed at points $x_i \in R^D$.

- **Model for observations**
  
  $y_i = f(x_i) + \text{“noise”}$ (Regression, eg. with positive noise )
  
  $y_i = \text{sign}[f(x_i) + \text{“noise”}]$ (Classification)

- **A priori information about “latent variable” (function $f$):**
  
  Realization of Gaussian random process with covariance $K(x, x')$. 
Ambiguities in local observation model for measuring wind velocity fields from satellites.

Solution: Model prior distribution of wind fields using a Gaussian process.
Example II: Code Division Multiple Access (CDMA)

- *K* users in mobile communication try to transmit message bits $S_1, \ldots S_K$ with $S_i \in \{-1, 1\}$ over single channel.

- **Modulation:** Multiply message with spreading code $x_k(n)$ for $n = 1, \ldots N_c$

- Received signals

$$ y(n) = \sum_{k=1}^{K} S_k x_k(n) + \sigma \varepsilon(n) $$

- **Inference:** Estimate $S_k$'s from the $y(n)$'s
  ($= \text{regression with binary variables}$).

(introduced to machine learning community by Toshiyuki Tanaka)
A canonical Class of Distributions

\[ P(S) = \frac{1}{Z} \prod_i \rho_i(S_i) \exp \left[ \sum_{i<j} S_i J_{ij} S_j \right] \]

\( \rho_i \) models local observations (likelihood) / or local constraints.

Normalization \( Z \) usually coincides with probability \( P(D) \) of observed data.
Problems with Inference

- Variables dependent $\to$ highdimensional integrals/sums.

- Exact inference impossible if random variables continuous (and non Gaussian).

- Laplace approximation for integrals impossible if integrand non differentiable.

- “Learning” of coupling matrix $J$ by EM-Algorithm (Maximum Likelihood) requires correlations $E[S_iS_j]$. 
Non-variational Approximations

- **Bethe approximation/Belief Propagation** (Yedidia, Freeman & Weiss):
  - "treelike" graphs.

- **TAP** - type of approximations: many neighbours, weak dependencies, Neighbourhood $\rightarrow$ Gaussian random influence.
Gibbs Free Energy

- Gives moments and $Z = P(D)$ simultaneously.

- Applicability of optimization methods

$$\Phi(m) = \min_Q \left\{ KL(Q||P) \mid E_Q[S_i] = m_i, E_Q[S_i^2] = M_i, i = 1, \ldots, N \right\} - \ln Z$$
TAP Approximation to Free Energy

Introduce tunable interaction strength $l$

$$P_l(S) = \frac{1}{Z} \prod_i \rho_i(S_i) \exp \left[ l \sum_{i<j} S_i J_{ij} S_j \right]$$

Exact result

$$\Phi_{l=1} = \Phi_{l=0} + \int_0^1 dl \frac{\partial \Phi_l}{\partial l} = \Phi_{l=0} - \frac{1}{2} \sum_{i,j} m_i J_{ij} m_j - \frac{1}{2} \int_0^1 dl \ \text{Tr}(C_l J)$$

with covariance $C_l$.

- TAP (Thouless, Anderson & Palmer) : Expand $\Phi_l$ to $O(l^2)$.

- Adaptive TAP (Opper & Winther): Gaussian approximation for $C_l$

$$C_l^g = (\Lambda_l - lJ)^{-1}$$
Properties of TAP Free Energy

- Free Energy has the form
  \[ \Phi^{\text{TAP}}(m, M) = \Phi_0(m, M) + \Phi^g(m, M) - \Phi^g_0(m, M) \]

  The \( \Phi \)'s are convex and correspond to
  - \( \Phi_0(m, M) \): true likelihood, no interactions.
  - \( \Phi^g(m, M) \): Gauss likelihood, full interactions.
  - \( \Phi^g_0(m, M) \): Gauss likelihood, no interactions.

- Minimizing hyperparameters of \( \Phi^{\text{TAP}} \) equal fixedpoints of approximate EM algorithm.
Relation to Cavity Approach

\[ \Phi_0 = \max_{\lambda^0, \gamma^0} \left\{ -\sum_i \ln Z_i^0(\gamma_i^0, \lambda_i^0) + \mathbf{m}^T \gamma^0 + \frac{1}{2} \mathbf{M}^T \lambda^0 \right\} \]

with

\[ Z_i(\gamma_i^0, \lambda_i^0) = \int dS \rho_i(S) \exp \left[ \gamma_i^0 S + \frac{1}{2} \lambda_i^0 S^2 \right] = \int dS \rho_i(S) E_z \left\{ \exp \left[ S(\gamma_i^0 + \sqrt{\lambda_i^0 z}) \right] \right\} \]

with \( z \) a standard normal Gaussian random variable.
Algorithm: Expectation Propagation (T. Minka)

Introduce effective Gaussian distribution having likelihood

\[
\prod_{i=1}^{N} \rho_i^g(S_i) = \prod_{i=1}^{N} e^{-\lambda_i S_i^2 + \gamma_i S_i}
\]

- → site \( i \). Replace Gaussian likelihood by true Likelihood.

  New Marginal

  \[
P_i(S) \propto P_i^g(S) \frac{\rho_i(S)}{\rho_i^g(S)} \rightarrow
\]

  Recompute \( E[S_i] \) and \( E[S_i^2] \)

- Recompute \( \lambda_i \) and \( \gamma_i \) → new site.
Exact Average case behaviour: Random J matrix ensembles, \( N \to \infty \)

Assume Orthogonal random matrix ensemble for \( J_N \) with

asymptotic scaling of generating function

\[
\frac{1}{N} \ln \left[ e^{\frac{1}{2} \text{Tr}(AJ_N)} \right] J \simeq \text{Tr} G(A/N)
\]

For \( N \to \infty \):

Average case properties (replica symmetry) of exact inference and ADATAP approximation agree (if single solution).
Application: Non-Gaussian Regression

\[ y = f(x) + \xi \text{ with positive noise } p(\xi) = \lambda e^{-\lambda x} I_{x>0} \text{: Estimate parameter } \lambda \text{ with } N = 1000. \]
Example: Estimation of Wind Fields

Likelihood  Monte Carlo prediction  ADATAP prediction
Results for Bayes optimal prediction $h_i = \text{artanh}(m_i)$: Exact/Mean Field and Exact/ADATAP.

$K = 8$ users and $N_c = 16$
Biterror Rate as a function of the number of users.

**SNR = 10dB** and Spreading factor $N_c = 20$
**Approximate analytical Bootstrap**

**Goal:** Estimate average case properties (eg test errors, uncertainty) of statistical predictor (eg SVM) **without** hold out test data.

**Bootstrap (Efron):** Generate **new pseudo training data** by resampling old training data with replacement.

Original training data: $D_0 = (z_1, z_2, z_3)$

Bootstrap samples: $D_1 = (z_1, z_1, z_2); D_2 = (z_1, z_2, z_2); D_3 = (z_3, z_3, z_3), \ldots$

**Problem:** Each sample requires time consuming retraining of predictor.

**Approximate analytical approach:** Average over samples with help of “replica trick”.
Supportvector Classifier (Vapnik)

SVM predicts \( y = \text{sign}[\hat{f}_{D_0}(x)] \) for \( x \in \mathbb{R}^d \),

with \( \hat{f}_{D_0}(x) = \sum_{j=1}^{N} y_j \alpha_i K(x, x_j) \) and \( K \) a positive definite kernel.

Setting \( S_i = \sum_{j=1}^{N} y_j \alpha_i K(x_i, x_j) \), the \( \alpha \)'s can be found from the convex optimization problem

\[
\text{Minimize} \quad \left\{ S^T K^{-1} S \right\} \quad \text{under the constraint} \quad S_i y_i \geq 1 , \quad i = 1, \ldots, N.
\]
Probabilistic formulation of Supportvector Machines

Define prior

\[ \mu[S] = \frac{1}{\sqrt{(2\pi)^N \beta^{-N}|K|}} \exp \left[ -\frac{\beta}{2} S^T K^{-1} S \right]. \]

and Pseudo-likelihood

\[ \prod_j P(y_j|S) = \prod_j \Theta(y_j S_j - 1) \]

where \( \Theta(u) = 1 \) for \( u > 0 \) and 0 otherwise.

For \( \beta \to \infty \), measure \( P[S|D] \propto \mu[S] P(D|S) \) concentrates at vector \( \hat{S} \) which solves SVM optimization problem.
Analytical Average using Replicas

Let $s_j = \#$ times data point $y_j$ appears in bootstrap sample $D$

\[
E_D[Z^n] = E_D \left[ \int \prod_{a=1}^{n} (dS^a \mu[S^a]) \prod_{j,a} P^{s_j}(y_j|S^a_j) \right] = \\
\int \prod_{a=1}^{n} (dS^a \mu[S^a]) \prod_{j=1}^{N} \exp \left[ \frac{S}{N} \prod_{a=1}^{n} P(y_j|S^a_j) \right]
\]

New intractable statistical model with coupled replicas!

Need approximate inference tools & limit $n \to 0$. 
Results: Classification & Regression

Compare TAP approximation theory / bootstrap simulation (= Sampling + Retraining)

Generalization error:

![Graph showing bootstrapped classification error vs bootstrap sample size for different datasets: Wisconsin, Pima, Sonar, Crabs, and Boston. Each dataset has a different number of test points, indicated as N=683, N=532, N=208, N=200, and N=506 respectively. The plots depict the decrease in error as the bootstrap sample size increases.]
SVM results cont’d

Uncertainty of SVM Prediction at test points

![Graph showing SVM results](image-url)
Regression

Distribution of predictor on training points
Outlook

• Systematic improvement

• Tractable substructures

• More complex dependencies (eg directed graphs)

• Fast algorithms & sparsity

• Combinatorial optimization problems, metastability

• Performance bounds?
Some worse Results