Algorithm A.10 Simple gradient ascent algorithm

```
    Procedure Gradient-Ascent (
      θ^1,    // Initial starting point
      f_obj,  // Function to be optimized
      δ       // Convergence threshold
    )
1     t ← 1
2     do
3       θ^{t+1} ← θ^t + η∇f_obj(θ^t)
4       t ← t + 1
5     while ||θ^t − θ^{t−1}|| > δ
6     return (θ^t)
```

search of algorithm A.5 (see appendix A.4.2). Using the Taylor expansion of a function, we know that, in the neighborhood of $\boldsymbol{\theta}^0$, the function can be approximated by the linear equation

$$f_{\text{obj}}(\boldsymbol{\theta}) \approx f_{\text{obj}}(\boldsymbol{\theta}^0) + (\boldsymbol{\theta} - \boldsymbol{\theta}^0)^T \nabla f_{\text{obj}}(\boldsymbol{\theta}^0).$$

Using basic properties of linear algebra, we can check that the slope of this linear function, that is, $\nabla f_{\text{obj}}(\boldsymbol{\theta}^0)$, points to the direction of the steepest ascent. This observation suggests that, if we take a step in the direction of the gradient, we increase the value of f_{obj}. This reasoning leads to the simple gradient ascent algorithm shown in algorithm A.10. Here, η is a constant that determines the *rate* of ascent at each iteration. Since the gradient ∇f_{obj} approaches 0 as we approach a maximum point, the procedure will converge if η is sufficiently small.

Note that, in order to apply gradient ascent, we need to be able to evaluate the function f_{obj} at different points, and also to evaluate its gradient. In several examples we encounter in this book, we can perform these calculations, although in some cases these are costly. Thus, a major objective is to reduce the number of points at which we evaluate f_{obj} or ∇f_{obj}.

The performance of gradient ascent depends on the choice of η. If η is too large, then the algorithm can "overshoot" the maximum in each iteration. For sufficiently small value of η, the gradient ascent algorithm will converge, but if η is too small, we will need many iterations to converge. Thus, one of the difficult points in applying this algorithm is deciding on the value of η. Indeed, in practice, one typically needs to begin with a large η, and decrease it over time; this approach leaves us with the problem of choosing an appropriate schedule for shrinking η.

A.5.2.2 Line Search

An alternative approach is to adaptively choose the step size η at each step. The intuition is that we choose a direction to climb and continue in that direction until we reach a point where we start to descend. In this procedure, at each point $\boldsymbol{\theta}^t$ in the search, we define a "line" in the direction of the gradient:

$$g(\eta) = \vec{\boldsymbol{\theta}^t} + \eta \nabla f_{\text{obj}}(\boldsymbol{\theta}^t).$$

line search

We now use a *line search* procedure to find the value of η that defines a (local) maximum of

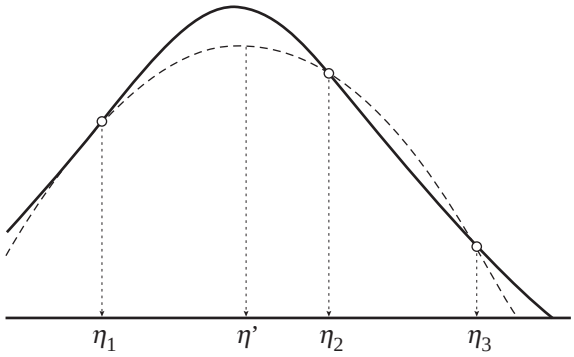


Figure A.2 Illustration of line search with Brent's method. The solid line shows a one-dimensional function. The three points, η_1, η_2, and η_3, bracket the maximum of this function. The dashed line shows the quadratic fit to these three points and the choice of η' proposed by Brent's method.

f_{obj} along the line; that is, we find:

$$\eta^t = \arg\max_{\eta} g(\eta).$$

We now take an η^t-sized step in the direction of the gradient; that is, we define:

$$\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t + \eta^t \nabla f_{\text{obj}}(\boldsymbol{\theta}^t).$$

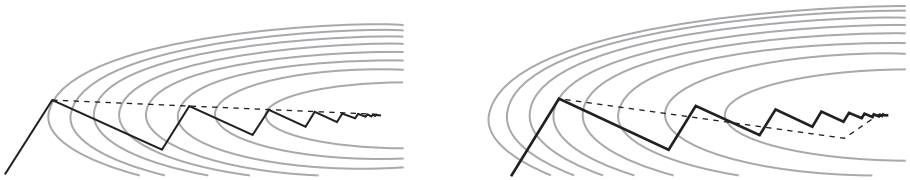
And the process repeats.

There are several methods for performing the line search. The basic idea is to find three points $\eta_1 < \eta_2 < \eta_3$ so that $f_{\text{obj}}(g(\eta_2))$ is larger than both $f_{\text{obj}}(g(\eta_1))$ and $f_{\text{obj}}(g(\eta_3))$. In this case, we know that there is at least one local maximum between η_1 and η_3, and we say that η_1, η_2 and η_3 *bracket* a maximum; see figure A.2 for an illustration. Once we have a method for finding a bracket, we can zoom in on the maximum. If we choose a point η' so that $\eta_1 < \eta' < \eta_2$ we can find a new, tighter, bracket. To see this, we consider the two possible cases. If $f_{\text{obj}}(g(\eta')) > f_{\text{obj}}(g(\eta_2))$, then η_1, η', η_2 bracket a maximum. Alternatively, if $f_{\text{obj}}(g(\eta')) \leq f_{\text{obj}}(g(\eta_2))$, then η', η_2, η_3 bracket a maximum. In both cases, the new bracket is smaller than the original one. Similar reasoning applies if we choose η' between η_2 and η_3.

The question is how to choose η'. One approach is to perform a binary search and choose $\eta' = (\eta_1 + \eta_3)/2$. This ensures that the size of the new bracket is half of the old one. A faster approach, known as *Brent's method*, fits a quadratic function based on the values of f_{obj} at the three points η_1, η_2, and η_3. We then choose η' to be the maximum point of this quadratic approximation. See figure A.2 for an illustration of this method.

A.5.2.3 Conjugate Gradient Ascent

Line search attempts to maximize the improvement along the direction defined by $\nabla f_{\text{obj}}(\boldsymbol{\theta}^t)$. This approach, however, often has undesired consequences on the convergence of the search. To understand the problem, we start by observing that $\nabla f_{\text{obj}}(\boldsymbol{\theta}^{t+1})$ must be *orthogonal* to



(a) (b)

Figure A.3 Two examples of the convergence problem with line search. The solid line shows the progression of gradient ascent with line search. The dashed line shows the progression of the conjugate gradient method: (a) a quadratic function $f_{\text{obj}}(x, y) = -(x^2 + 10y^2)$; (b) its exponential $f_{\text{obj}}(x, y) = \exp\{-(x^2 + 10y^2)\}$. In both cases, the two search procedures start from the same initial point (bottom left of the figure), and diverge after the first line search.

$\nabla f_{\text{obj}}(\boldsymbol{\theta}^t)$. To see why, observe that $\boldsymbol{\theta}^{t+1}$ was chosen to be a local maximum along the $\nabla f_{\text{obj}}(\boldsymbol{\theta}^t)$ direction. Thus, the gradient of f_{obj} at $\boldsymbol{\theta}^{t+1}$ must be 0 in this direction. This implies that the two consecutive gradient vectors are orthogonal. As a consequence, the progress of the gradient ascent will be in a zigzag line. As the procedure approaches a maximum point, the size of each step becomes smaller, and the progress slows down. See figure A.3 for an illustration of this phenomenon.

A possible solution is to "remember" past directions of search and to bias the new direction to be a combination of the gradient at the current point and the direction implied by previous steps. This intuitive idea can be developed into a variety of algorithms. It turns out, however, that one variant of this algorithm can be shown to be optimal for finding the maximum of quadratic functions. Since, by the Taylor expansion, all functions are approximately quadratic in the neighborhood of a maximum, it follows that the final steps of the algorithm will converge to a maximum relatively quickly.

conjugate gradient ascent

The algorithm, known as *conjugate gradient ascent*, is shown in algorithm A.11. The vector $\boldsymbol{h}^t$ is the "corrected" direction for search. It combines the gradient $\boldsymbol{g}^t$ with the previous direction of search $\boldsymbol{h}^{t-1}$. The effect of previous search directions on the new one depends on the relative sizes of the gradients.

If our function f_{obj} is a quadratic function, the conjugate gradient ascent procedure is guaranteed to converge in n steps, where n is the dimension of the space. Indeed, in figure A.3a we see that the conjugate method converges in two steps. When the function is not quadratic, conjugate gradient ascent might require more steps, but is still much faster than standard gradient ascent. For example, in figure A.3b, it converges in four steps (the last step is too small to be visible in the figure).

Finallly, we note that **gradient ascent is the continuous analogue of the local hill-climbing approaches described in section A.4.2. As such, it is susceptible to the same issues of local maxima and plateaus. The approaches used to address these issues in this setting are similar to those outlined in the discrete case.**

Algorithm A.11 Conjugate gradient ascent

Procedure Conjugate-Gradient-Ascent (
  $\boldsymbol{\theta}^1$, // Initial starting point
  f_{obj}, // Function to be optimized
  δ // Convergence threshold
)
  $t \leftarrow 1$
  $\boldsymbol{g}^0 \leftarrow \mathbf{1}$
  $\boldsymbol{h}^0 \leftarrow \mathbf{0}$
  do
    $\boldsymbol{g}^t \leftarrow \nabla f_{\text{obj}}(\boldsymbol{\theta}^t)$
    $\gamma^t \leftarrow \frac{(\boldsymbol{g}^t - \boldsymbol{g}^{t-1})^T \boldsymbol{g}^t}{(\boldsymbol{g}^{t-1})^T \boldsymbol{g}^{t-1}}$
    $\boldsymbol{h}^t \leftarrow \boldsymbol{g}^t + \gamma^t \boldsymbol{h}^{t-1}$
    Choose η^t by line search along the line $\boldsymbol{\theta}_t + \eta \boldsymbol{h}^t$
    $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t + \eta^t \boldsymbol{h}^t$
    $t \leftarrow t + 1$
  while $\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\| > \delta$
  return $(\boldsymbol{\theta}^t)$

A.5.3 Constrained Optimization

In appendix A.5.1, we considered the problem of optimizing a continuous function over its entire domain (see also appendix A.5.2). In many cases, however, we have certain constraints that the desired solution must satisfy. Thus, we have to optimize the function within a constrained space. We now review some basic methods that address this problem of *constrained optimization.*

constrained optimization

Example A.5

Suppose we want to find the maximum entropy distribution over a variable X, with $Val(X) = \{x^1, \ldots, x^K\}$. Consider the entropy of X:

$$\mathbf{H}(X) = -\sum_{k=1}^{K} P(x^k) \log P(x^k).$$

We can maximize this function using the gradient method by treating each $P(x^k)$ as a separate parameter θ_k. We compute the gradient of $\mathbf{H}_P(X)$ with respect to each of these parameters:

$$\frac{\partial}{\partial \theta_k} \mathbf{H}(X) = -\log(\theta_k) - 1.$$

Setting this partial derivative to 0, we get that $\log(\theta_k) = -1$, and thus $\theta_k = 1/2$. This solution seems fine until we realize that the numbers do not sum up to 1, and hence our solution does not *define a probability distribution!*

The flaw in our analysis is that we want to maximize the entropy subject to a constraint *on the parameters, namely, $\sum_k \theta_k = 1$. In addition, we also remember that we need to require that $\theta_k \geq 0$. In this case we see that the gradient drives the solution away from from 0 ($-\log(\theta_k) \to \infty$ as $\theta_k \to 0$), and thus we do not need to enforce this constraint actively.* ■

equality constraint

Problems of this type appear in many settings, where we are interested in maximizing a function $\boldsymbol{f}$ under a set of *equality constraints*. This problem is posed as follows:

$$\begin{array}{lll} \textbf{Find} & \boldsymbol{\theta} & \\ \textbf{maximizing} & \boldsymbol{f}(\boldsymbol{\theta}) & \\ \textbf{subject to} & & \\ & c_1(\boldsymbol{\theta}) = 0 & \\ & \dots & \\ & c_m(\boldsymbol{\theta}) = 0. & \end{array} \tag{A.5}$$

Note that any equality constraint (such as the one in our example above) can be rephrased as constraining a function c to 0. Formally, we are interested in the behavior of $\boldsymbol{f}$ in the region of points that satisfies all the constraints

$$\mathcal{C} = \{\boldsymbol{\theta} : \forall j = 1, \dots, n,\ c_j(\boldsymbol{\theta}) = 0\}.$$

To define our goal, remember that we want to find a maxima point within $\mathcal{C}$. Since $\mathcal{C}$ is a constrained "surface" we need to adopt the basic definition of maxima (and similarly minima, stationary point, etc.) to this situation. We can define local maxima in two ways. The first definition is in term of neighborhood. We define the ϵ-neighborhood of $\boldsymbol{\theta}$ in $\mathcal{C}$ to be all the points $\theta' \in \mathcal{C}$ such that $\|\theta - \theta'\|_2 < \epsilon$. We then say that θ is a local maxima in $\mathcal{C}$ if there is an $\epsilon > 0$ such that $\boldsymbol{f}(\theta) > \boldsymbol{f}(\theta')$ for all θ' in its ϵ-neighborhood. An alternative definition that will be easier for the following is in terms of derivatives. Recall that a stationary point (local maximum, local minimum, or a saddle point) of a function if the derivative is 0. In the constraint case we have a similar definition, but we must ensure that the derivatives are ones that do not take us outside the constrained surface. Stated differently, if we consider a derivative in the direction $\boldsymbol{\delta}$, we want to ensure that the constraints remain 0 if we take a small step in direction $\boldsymbol{\delta}$. Formally, this means that the derivative has to be *tangent* to each constraint c_i, that is $\boldsymbol{\delta}^T \nabla c_i(\boldsymbol{\theta}) = 0$.

Lagrange multipliers

A general approach to solving such constrained optimization problems is the method of *Lagrange multipliers*. We define a new function, called the *Lagrangian*, of $\boldsymbol{\theta}$ and of a new vector of parameters $\boldsymbol{\lambda} = \langle \lambda_1, \dots, \lambda_m \rangle$

$$\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \boldsymbol{f}(\boldsymbol{\theta}) - \sum_{j=1}^{m} \lambda_j c_j(\boldsymbol{\theta}).$$

Theorem A.7 *If $\langle \boldsymbol{\theta}, \boldsymbol{\lambda} \rangle$ is a stationary point of the Lagrangian $\mathcal{J}$, then $\boldsymbol{\theta}$ is a stationary point of $\boldsymbol{f}$ subject to the constraints $c_1(\boldsymbol{\theta}) = 0, \dots, c_m(\boldsymbol{\theta}) = 0$.*

Proof We briefly outline the proof. A formal proof requires the use of more careful tools from functional analysis.

We start by showing that $\boldsymbol{\theta}$ satisfies the constraints. Since $\langle \boldsymbol{\theta}, \boldsymbol{\lambda} \rangle$ is a stationary point of $\mathcal{J}$, we have that for each j

$$\frac{\partial}{\partial \lambda_j} \mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = -c_j(\boldsymbol{\theta}).$$

Thus, at stationary points of $\mathcal{J}$, the constraint $c_j(\boldsymbol{\theta}) = 0$ must be satisfied.

Now consider $\nabla \boldsymbol{f}(\boldsymbol{\theta})$. For each component θ_i of $\boldsymbol{\theta}$, we have that

$$0 = \frac{\partial}{\partial \theta_i} \mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{\partial}{\partial \theta_i} \boldsymbol{f}(\boldsymbol{\theta}) - \sum_j \lambda_j \frac{\partial}{\partial \theta_i} c_j(\boldsymbol{\theta}).$$

Thus,

$$\nabla \boldsymbol{f}(\boldsymbol{\theta}) = \sum_j \lambda_j \nabla c_j(\boldsymbol{\theta}). \tag{A.6}$$

In other words, the gradient of $\boldsymbol{f}$ is a linear combination of the gradients of c_j.

We now use this property to prove that $\boldsymbol{\theta}$ is a stationary point of $\boldsymbol{f}$ when constrained to region $\mathcal{C}$. Consider a direction $\boldsymbol{\delta}$ that is tangent to the region $\mathcal{C}$ at $\boldsymbol{\theta}$. As $\boldsymbol{\delta}$ is tangent to $\mathcal{C}$, we expect that moving infinitesimally in this direction will maintain the constraint that c_j is 0; that is, c_j should not change its value when we move in this direction. More formally, the derivative of c_j in the direction $\boldsymbol{\delta}$ is 0. The derivative of c_j in a direction $\boldsymbol{\delta}$ is $\boldsymbol{\delta}^T \nabla c_j$. Thus, if $\boldsymbol{\delta}$ is tangent to $\mathcal{C}$, we have

$$\boldsymbol{\delta}^T \nabla c_j(\boldsymbol{\theta}) = 0$$

for all j. Using equation (A.6), we get

$$\boldsymbol{\delta}^T \nabla \boldsymbol{f}(\boldsymbol{\theta}) = \sum_j \lambda_j \boldsymbol{\delta}^T \nabla c_j(\boldsymbol{\theta}) = 0.$$

Thus, the derivative of $\boldsymbol{f}$ in a direction that is tangent to $\mathcal{C}$ is 0. This implies that when moving away from $\boldsymbol{\theta}$ within the allowed region $\mathcal{C}$ the value of $\boldsymbol{f}$ has 0 derivative. Thus, $\boldsymbol{\theta}$ is a stationary point of $\boldsymbol{f}$ when restricted to $\mathcal{C}$. ■

We also have the converse property: If $\boldsymbol{f}$ satisfies some regularity conditions, then for every stationary point of $\boldsymbol{f}$ in $\mathcal{C}$ there is a choice of $\boldsymbol{\lambda}$ so that $\langle \boldsymbol{\theta}, \boldsymbol{\lambda} \rangle$ is a stationary point of $\mathcal{J}$.

We see that the Lagrangian construction allows us to solve constrained optimization problems using tools for unconstrained optimization. We note that a local maximum of $\boldsymbol{f}$ always corresponds to a stationary point of $\mathcal{J}$, but this stationary point is not necessarily a local maximum of $\mathcal{J}$. If, however, we restrict attention to nonnegative constraint functions c, then a local maximum of $\boldsymbol{f}$ must correspond to a local maximum of $\mathcal{J}$.

We now consider two examples of using this technique.

Example A.6

Let us return to example A.5. In order to find the maximum entropy distribution over X, we need to solve the Lagrangian

$$\mathcal{J} = -\sum_k \theta_k \log \theta_k - \lambda \left(\sum_k \theta_k - 1 \right).$$

Setting $\nabla \mathcal{J} = 0$ implies the following system of equations:

$$\begin{aligned} 0 &= -\log \theta_1 - 1 - \lambda \\ &\cdots \\ 0 &= -\log \theta_K - 1 - \lambda \\ 0 &= \sum_k \theta_k - 1. \end{aligned}$$

Each of the first K equations can be rewritten as $\theta_k = 2^{-1-\lambda}$. Plugging this term into the last equation, we get that $\lambda = \log(K) - 1$, and thus $P(x^k) = 1/K$. We conclude that we achieve maximum entropy with the uniform distribution. ■

To see an example with more than one constraint, consider the following problem.

Example A.7

M-projection

Suppose we have a distribution $P(X, Y)$ over two random variables, and we want to find the closest distribution $Q(X, Y)$ in which X is independent of Y. As we discussed in section 8.5, this process is called M-projection *(see definition 8.4). Since X and Y are independent in Q, we must have that $Q(X, Y) = Q(X)Q(Y)$. Thus, we are searching for parameters $\theta_x = Q(x)$ and $\theta_y = Q(y)$ for different values $x \in Val(X)$ and $y \in Val(Y)$.*

Formally, we want to solve the following problem:

Find $\{\theta_x : x \in Val(X)\}$ and $\{\theta_y : y \in Val(y)\}$ that minimize

$$\boldsymbol{D}(P(X,Y) \| Q(X)Q(Y)) = \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{\theta_x \theta_y},$$

subject to the constraints

$$\begin{aligned} 0 &= \sum_x \theta_x - 1 \\ 0 &= \sum_y \theta_y - 1. \end{aligned}$$

We define the Lagrangian

$$\mathcal{J} = \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{\theta_x \theta_y} - \lambda_x \left(\sum_x \theta_x - 1 \right) - \lambda_y \left(\sum_y \theta_y - 1 \right).$$

To simplify the computation of derivatives, we notice that

$$\log \frac{P(x,y)}{\theta_x \theta_y} = \log P(x,y) - \log \theta_x - \log \theta_y.$$

Using this simplification, we can compute the derivative with respect to the probability of a particular value of X, say θ_{x^k}. We note that this parameter appears only when the value of x in the summation equals x^k. Thus,

$$\frac{\partial}{\partial \theta_{x^k}} \mathcal{J} = -\sum_y \frac{P(x^k, y)}{\theta_{x^k}} - \lambda_x.$$

Equating this derivative to 0*, we get*

$$\theta_{x^k} = -\frac{\sum_y P(x^k, y)}{\lambda_x} = -\frac{P(x^k)}{\lambda_x}.$$

To solve for the value of λ_x*, we use the first constraint, and get that*

$$1 = \sum_x \theta_x = -\sum_x \frac{P(x)}{\lambda_x}.$$

■

Thus, we get that $\lambda_x = -\sum_x P(x)$*. Thus, we can conclude that* $\lambda_x = -1$*, and consequently that* $\theta_x = P(x)$*. An analogous reasoning shows that* $\theta_y = P(y)$*.*

This solution is very natural. The closest distribution to $P(X, Y)$ *in which* X *and* Y *are independent is* $Q(X, Y) = P(X)P(Y)$*. This distribution preserves the marginal distributions of both* X *and* Y*, but loses all information about their joint behavior.*

A.5.4 Convex Duality

convex duality

The concept of *convex duality* plays a central role in optimization theory. We briefly review the main results here for equality-constrained optimization problems with nonnegativity constraints (although the theory extends quite naturally to the case of general inequality constraints).

In appendix A.5.3, we considered an optimization problem of maximizing $f(\boldsymbol{\theta})$ subject to certain constraints, which we now call the *primal problem*. We showed how to formulate a Lagrangian $\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\lambda})$, and proved that if $\langle \boldsymbol{\theta}, \boldsymbol{\lambda} \rangle$ is a stationary point of $\mathcal{J}$ then $\boldsymbol{\theta}$ is a stationary point of the objective function f that we are trying to maximize.

We can extend this idea further and define the *dual function* $g(\boldsymbol{\lambda})$ as

$$g(\boldsymbol{\lambda}) = \sup_{\boldsymbol{\theta} \geq 0} \mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\lambda}).$$

That is, the dual function $g(\boldsymbol{\lambda})$, is the *supremum*, or maximum, over the parameters $\boldsymbol{\theta}$ for a given $\boldsymbol{\lambda}$. In general, we allow the dual function to take the value ∞ when $\mathcal{J}$ is unbounded above (which can occur when the primal constraints are unsatisfied), and refer to the points $\boldsymbol{\lambda}$ at which this happens as *dual infeasible*.

Example A.8

Let us return to example A.6, where our task is to find the distribution $P(X)$ *of maximum entropy. Now, however, we also want the distribution to satisfy the constraint that* $\mathbf{E}_P[X] = \mu$*. Treating each* $P(X = k)$ *as a separate parameter* θ_k*, we can write our problem formally as:*

Constrained-Entropy*:*

Find	P
maximizing	$H_P(X)$
subject to	
	$\sum_{k=1}^K k\theta_k = \mu$
	$\sum_{k=1}^K \theta_k = 1$
	$\theta_k \geq 0 \qquad \forall k = 1, \ldots, K$

(A.7)

Lagrange multipliers

Introducing Lagrange multipliers *for each of the constraints we can write*

$$\mathcal{J}(\boldsymbol{\theta}, \lambda, \nu) = -\sum_{k=1}^{K} \theta_k \log \theta_k - \lambda \left(\sum_{k=1}^{K} k\theta_k - \mu \right) - \nu \left(\sum_{k=1}^{K} \theta_k - 1 \right).$$

Maximizing over $\boldsymbol{\theta}$ *for each* $\langle \lambda, \nu \rangle$ *we get the dual function*

$$\begin{aligned} \boldsymbol{g}(\lambda, \nu) &= \sup_{\boldsymbol{\theta} \geq 0} \mathcal{J}(\boldsymbol{\theta}, \lambda, \nu) \\ &= \lambda\mu + \nu + e^{-\nu-1} \sum_{k} e^{-k\lambda}. \end{aligned}$$

Thus, the convex dual (to be minimized) is $\lambda\mu + \nu + e^{-\nu-1} \sum_k e^{-k\lambda}$. *We can minimize over* ν *analytically by taking derivatives and setting them equal to zero, giving* $\nu = \log \boldsymbol{g}(\sum_k e^{-k\lambda}) - 1$. *Substituting into* $\boldsymbol{g}$, *we arrive at the dual optimization problem*

$$\textit{minimize} \quad \lambda\mu + \log \left(\textstyle\sum_{k=1}^{K} e^{-k\lambda} \right).$$

This form of optimization problem is known as a geometric program. *The convexity of the objective function can be easily verified by taking second derivatives. Taking the first derivative and setting it to zero provides some insight into the solution to the problem:*

$$\frac{\sum_{k=1}^{K} k e^{-k\lambda}}{\sum_{k=1}^{K} e^{-k\lambda}} = \mu,$$

indicating that the solution has $\theta_k \propto \alpha^k$ *for some fixed* α. ■

Importantly, as we can see in this example, the dual function is a pointwise maximization over a family of linear functions (of the dual variables). Thus, the dual function is always convex even when the primal objective function $\boldsymbol{f}$ is not.

One of the most important results in optimization theory is that the dual function gives an upper bound on the optimal value of the optimization problem; that is, for any primal feasible point $\boldsymbol{\theta}$ and any dual feasible point $\boldsymbol{\lambda}$, we have $\boldsymbol{g}(\boldsymbol{\lambda}) \geq f_{\text{obj}}(\boldsymbol{\theta})$. This leads directly to the property of *weak duality*, which states that the minimum value of the dual function is at least as large as the maximum value of the primal problem; that is,

$$\boldsymbol{g}(\boldsymbol{\lambda}^\star) = \inf_{\boldsymbol{\lambda}} \boldsymbol{g}(\boldsymbol{\lambda}) \geq \boldsymbol{f}(\boldsymbol{\theta}^\star).$$

The difference $\boldsymbol{f}(\boldsymbol{\theta}^\star) - \boldsymbol{g}(\boldsymbol{\lambda}^\star)$ is known as the *duality gap*. Under certain conditions the duality gap is zero, that is, $\boldsymbol{f}(\boldsymbol{\theta}^\star) = \boldsymbol{g}(\boldsymbol{\lambda}^\star)$, in which case we have *strong duality*. Thus, duality can be used to provide a *certificate* of optimality. That is, if we can show that $\boldsymbol{g}(\boldsymbol{\lambda}) = \boldsymbol{f}(\boldsymbol{\theta})$ for some value of $\langle \boldsymbol{\theta}, \boldsymbol{\lambda} \rangle$, then we know that $\boldsymbol{f}(\boldsymbol{\theta})$ is optimal.

The concept of a dual function plays an important role in optimization. In a number of situations, the dual objective function is easier to optimize than the primal. Moreover, there are methods that solve the primal and dual together, using the fact that each bounds the other to improve the search for an optimal solution.

Bibliography

Abbeel, P., D. Koller, and A. Ng (2006, August). Learning factor graphs in polynomial time & sample complexity. *Journal of Machine Learning Research 7*, 1743–1788.

Ackley, D., G. Hinton, and T. Sejnowski (1985). A learning algorithm for Boltzmann machines. *Cognitive Science 9*, 147–169.

Aji, S. M. and R. J. McEliece (2000). The generalized distributive law. *IEEE Trans. Information Theory 46*, 325–343.

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automatic Control 19*, 716–723.

Akashi, H. and H. Kumamoto (1977). Random sampling approach to state estimation in switching environments. *Automatica 13*, 429–434.

Allen, D. and A. Darwiche (2003a). New advances in inference by recursive conditioning. In *Proc. 19th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 2–10.

Allen, D. and A. Darwiche (2003b). Optimal time–space tradeoff in probabilistic inference. In *Proc. 18th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 969–975.

Altun, Y., I. Tsochantaridis, and T. Hofmann (2003). Hidden Markov support vector machines. In *Proc. 20th International Conference on Machine Learning (ICML)*.

Andersen, S., K. Olesen, F. Jensen, and F. Jensen (1989). HUGIN—a shell for building Bayesian belief universes for expert systems. In *Proc. 11th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1080–1085.

Anderson, N. (1974). Information integration theory: A brief survey. In *Contemporary developments in Mathematical Psychology*, Volume 2, pp. 236–305. San Francisco, California: W.H. Freeman and Company.

Anderson, N. (1976). How functional measurement can yield validated interval scales of mental quantities. *Journal of Applied Psychology 61*(6), 677–692.

Andreassen, S., F. Jensen, S. Andersen, B. Falck, U. Kjærulff, M. Woldbye, A. R. Sørensen, A. Rosenfalck, and F. Jensen (1989). MUNIN — an expert EMG assistant. In J. E. Desmedt (Ed.), *Computer-Aided Electromyography and Expert Systems*, Chapter 21. Amsterdam: Elsevier Science Publishers.

Anguelov, D., D. Koller, P. Srinivasan, S. Thrun, H.-C. Pang, and J. Davis (2004). The correlated correspondence algorithm for unsupervised registration of nonrigid surfaces. In *Proc. 18th Conference on Neural Information Processing Systems (NIPS)*.

Arnauld, A. and P. Nicole (1662). Port-royal logic.

Arnborg, S. (1985). Efficient algorithms for combinatorial problems on graphs with bounded, decomposability—a survey. *BIT 25*(1), 2–23.

Arnborg, S., D. Corneil, and A. Proskurowski (1987). Complexity of finding embeddings in a k-tree. *SIAM J. Algebraic Discrete Methods 8*(2), 277–284.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 21–30.

Avriel, M. (2003). *Nonlinear Programming: Analysis and Methods*. Dover Publishing.

Bacchus, F. and A. Grove (1995). Graphical models for preference and utility. In *Proc. UAI–95*, pp. 3–10.

Bach, F. and M. Jordan (2001). Thin junction trees. In *Proc. 15th Conference on Neural Information Processing Systems (NIPS)*.

Balke, A. and J. Pearl (1994a). Counterfactual probabilities: Computational methods, bounds and applications. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 46–54.

Balke, A. and J. Pearl (1994b). Probabilistic evaluation of counterfactual queries. In *Proc. 10th Conference on Artificial Intelligence (AAAI)*, pp. 230–237.

Bar-Shalom, Y. (Ed.) (1992). *Multitarget multisensor tracking: Advanced applications*. Norwood, Massachusetts: Artech House.

Bar-Shalom, Y. and T. Fortmann (1988). *Tracking and Data Association*. New York: Academic Press.

Bar-Shalom, Y., X. Li, and T. Kirubarajan (2001). *Estimation with Application to Tracking and Navigation*. John Wiley and Sons.

Barash, Y. and N. Friedman (2002). Context-specific Bayesian clustering for gene expression data. *Journal of Computational Biology 9*, 169–191.

Barber, D. and W. Wiegerinck (1998). Tractable variational structures for approximating graphical models. In *Proc. 12th Conference on Neural Information Processing Systems (NIPS)*, pp. 183–189.

Barbu, A. and S. Zhu (2005). Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities. *IEEE Trans. on Pattern Analysis and Machine Intelligence 27*(8), 1239–1253.

Barnard, S. (1989). Stochastic stero matching over scale. *International Journal of Computer Vision 3*, 17–32.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley.

Barron, A., J. Rissanen, and B. Yu (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory 44*(6), 2743–2760.

Bartlett, M. (1935). Contingency table interactions. *Journal of the Royal Statistical Society, Series B 2*, 248–252.

Bauer, E., D. Koller, and Y. Singer (1997). Update rules for parameter estimation in Bayesian networks. In *Proc. 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 3–13.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London 53*, 370–418.

Beal, M. and Z. Ghahramani (2006). Variational Bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis 1*, 793–832.

Becker, A., R. Bar-Yehuda, and D. Geiger (1999). Random algorithms for the loop cutset problem. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 49–56.

Becker, A. and D. Geiger (1994). The loop cutset problem. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 60–68.

Becker, A. and D. Geiger (2001). A sufficiently fast algorithm for finding close to optimal clique

trees. *Artificial Intelligence 125*(1–2), 3–17.

Becker, A., D. Geiger, and C. Meek (2000). Perfect tree-like Markovian distributions. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 19–23.

Becker, A., D. Geiger, and A. Schäffer (1998). Automatic selection of loop breakers for genetic linkage analysis. *Human Heredity 48*, 49–60.

Beeri, C., R. Fagin, D. Maier, and M. Yannakakis (1983). On the desirability of acyclic database schemes. *Journal of the Association for Computing Machinery 30*(3), 479–513.

Beinlich, L., H. Suermondt, R. Chavez, and G. Cooper (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the Second European Conference on Artificial Intelligence in Medicine*, pp. 247–256. Springer Verlag.

Bell, D. (1982). egret in decision making under uncertainty. *Operations Research 30*, 961–981.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton, New Jersey: Princeton University Press.

Ben-Tal, A. and A. Charnes (1979). A dual optimization framework for some problems of information theory and statistics. *Problems of Control and Information Theory 8*, 387–401.

Bentham, J. (1789). An introduction to the principles of morals and legislation.

Berger, A., S. Della-Pietra, and V. Della-Pietra (1996). A maximum entropy approach to natural language processing. *Computational Linguistics 16*(2).

Bernardo, J. and A. Smith (1994). *Bayesian Theory*. New York: John Wiley and Sons.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis (exposition of a new theory on the measurement of risk). English Translation by L. Sommer, *Econometrica*, 22:23–36, 1954.

Berrou, C., A. Glavieux, and P. Thitimajshima (1993). Near Shannon limit error-correcting coding: Turbo codes. In *Proc. International Conference on Communications*, pp. 1064–1070.

Bertelé, U. and F. Brioschi (1972). *Nonserial Dynamic Programming*. New York: Academic Press.

Bertsekas, D. (1999). *Nonlinear Programming* (2nd ed.). Athena Scientific.

Bertsekas, D. P. and J. N. Tsitsiklis (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Besag, J. (1977a). Efficiency of pseudo-likelihood estimation for simple Gaussian fields. *Biometrika 64*(3), 616–618.

Besag, J. (1977b). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B 36*, 192–236.

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society, Series B 48*, 259–302.

Bethe, H. A. (1935). Statistical theory of superlattices. *in Proceedings of the Royal Society of London A*, 552.

Bidyuk, B. and R. Dechter (2007). Cutset sampling for bayesian networks. *Journal of Artificial Intelligence Research 28*, 1–48.

Bilmes, J. and C. Bartels (2003). On triangulating dynamic graphical models. In *Proc. 19th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Bilmes, J. and C. Bartels (2005, September). Graphical model architectures for speech recognition. *IEEE Signal Processing Magazine 22*(5), 89–100.

Binder, J., D. Koller, S. Russell, and K. Kanazawa (1997). Adaptive probabilistic networks with hidden variables. *Machine Learning 29*, 213–244.

Binder, J., K. Murphy, and S. Russell (1997). Space-efficient inference in dynamic probabilistic networks. In *Proc. 15th International Joint Conference on Artificial Intelligence (IJCAI)*.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics

(M. Jordan, J. Kleinberg, and B. Schökopf, editors). New York: Springer-Verlag.

Bishop, C., N. Lawrence, T. Jaakkola, and M. Jordan (1997). Approximating posterior distributions in belief networks using mixtures. In *Proc. 11th Conference on Neural Information Processing Systems (NIPS)*.

Blalock, Jr., H. (1971). *Causal Models in the Social Sciences*. Chicago, Illinois: Aldine-Atheson.

Blum, B., C. Shelton, and D. Koller (2006). A continuation method for nash equilibria in structured games. *Journal of Artificial Intelligence Resarch 25*, 457–502.

Bodlaender, H., A. Koster, F. van den Eijkhof, and L. van der Gaag (2001). Pre-processing for triangulation of probabilistic networks. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 32–39.

Boros, E. and P. Hammer (2002). Pseudo-Boolean optimization. *Discrete Applied Mathematics 123*(1-3).

Bouckaert, R. (1993). Probabilistic network construction using the minimum description length principle. In *Proc. European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pp. 41–48.

Boutilier, C. (2002). A POMDP formulation of preference elicitation problems. In *Proc. 18th Conference on Artificial Intelligence (AAAI)*, pp. 239–46.

Boutilier, C., F. Bacchus, and R. Brafman (2001). UCP-Networks: A directed graphical representation of conditional utilities. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 56–64.

Boutilier, C., T. Dean, and S. Hanks (1999). Decision theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research 11*, 1 – 94.

Boutilier, C., R. Dearden, and M. Goldszmidt (1989). Exploiting structure in policy construction. In *Proc. 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1104–1111.

Boutilier, C., R. Dearden, and M. Goldszmidt (2000). Stochastic dynamic programming with factored representations. *Artificial Intelligence 121*(1), 49–107.

Boutilier, C., N. Friedman, M. Goldszmidt, and D. Koller (1996). Context-specific independence in Bayesian networks. In *Proc. 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 115–123.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Boyen, X., N. Friedman, and D. Koller (1999). Discovering the hidden structure of complex dynamic systems. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 91–100.

Boyen, X. and D. Koller (1998a). Approximate learning of dynamic models. In *Proc. 12th Conference on Neural Information Processing Systems (NIPS)*.

Boyen, X. and D. Koller (1998b). Tractable inference for complex stochastic processes. In *Proc. 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 33–42.

Boyen, X. and D. Koller (1999). Exploiting the architecture of dynamic systems. In *Proc. 15th Conference on Artificial Intelligence (AAAI)*.

Boykov, Y., O. Veksler, and R. Zabih (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*(11), 1222–1239.

Braziunas, D. and C. Boutilier (2005). Local utility elicitation in GAI models. In *Proc. 21st Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 42–49.

Breese, J. and D. Heckerman (1996). Decision-theoretic troubleshooting: A framework for repair and experiment. In *Proc. 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp.

124–132.

Breese, J., D. Heckerman, and C. Kadie (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 43–52.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees*. Monterey,CA: Wadsworth & Brooks.

Buchanan, B. and E. Shortliffe (Eds.) (1984). *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley.

Bui, H., S. Venkatesh, and G. West (2001). Tracking and surveillance in wide-area spatial environments using the Abstract Hidden Markov Model. *International Journal of Pattern Recognition and Artificial Intelligence*.

Buntine, W. (1991). Theory refinement on Bayesian networks. In *Proc. 7th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 52–60.

Buntine, W. (1993). Learning classification trees. In D. J. Hand (Ed.), *Artificial Intelligence Frontiers in Statistics*, Number III in AI and Statistics. Chapman & Hall.

Buntine, W. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research 2*, 159–225.

Buntine, W. (1996). A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering 8*, 195–210.

Caffo, B., W. Jank, and G. Jones (2005). Ascent-based Monte Carlo Expectation-Maximization. *Journal of the Royal Statistical Society, Series B*.

Cannings, C., E. A. Thompson, and H. H. Skolnick (1976). The recursive derivation of likelihoods on complex pedigrees. *Advances in Applied Probability 8*(4), 622–625.

Cannings, C., E. A. Thompson, and M. H. Skolnick (1978). Probability functions on complex pedigrees. *Advances in Applied Probability 10*(1), 26–61.

Cano, J., L.D., Hernández, and S. Moral (2006). Importance sampling algorithms for the propagation of probabilities in belief networks. *International Journal of Approximate Reasoning 15*(1), 77–92.

Carreira-Perpignan, M. and G. Hinton (2005). On contrastive divergence learning. In *Proc. 11thWorkshop on Artificial Intelligence and Statistics*.

Casella, G. and R. Berger (1990). *Statistical Inference*. Wadsworth.

Castillo, E., J. Gutiérrez, and A. Hadi (1997a). *Expert Systems and Probabilistic Network Models*. New York: Springer-Verlag.

Castillo, E., J. Gutiérrez, and A. Hadi (1997b). Sensitivity analysis in discrete Bayesian networks. *IEEE Transactions on Systems, Man and Cybernetics 27*, 412–23.

Chajewska, U. (2002). *Acting Rationally with Incomplete Utility Information*. Ph.D. thesis, Stanford University.

Chajewska, U. and D. Koller (2000). Utilities as random variables: Density estimation and structure discovery. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 63–71.

Chajewska, U., D. Koller, and R. Parr (2000). Making rational decisions using adaptive utility elicitation. In *Proc. 16th Conference on Artificial Intelligence (AAAI)*, pp. 363–369.

Chan, H. and A. Darwiche (2002). When do numbers really matter? *Journal of Artificial Intelligence Research 17*, 265–287.

Chávez, T. and M. Henrion (1994). Efficient estimation of the value of information in Monte Carlo

models. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 119–127.

Cheeseman, P., J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman (1988). Autoclass: a Bayesian classification system. In *Proc. 5th International Conference on Machine Learning (ICML)*.

Cheeseman, P., M. Self, J. Kelly, and J. Stutz (1988). Bayesian classification. In *Proc. 4th Conference on Artificial Intelligence (AAAI)*, Volume 2, pp. 607–611.

Cheeseman, P. and J. Stutz (1995). Bayesian classification (AutoClass): Theory and results. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*. AAAI Press.

Chen, L., M. Wainwright, M. Cetin, and A. Willsky (2003). Multitarget-multisensor data association using the tree-reweighted max-product algorithm. In *Proceedings SPIE Aerosense Conference*, Orlando, Florida.

Chen, R. and S. Liu (2000). Mixture Kalman filters. *Journal of the Royal Statistical Society, Series B*.

Cheng, J. and M. Druzdzel (2000). AIS-BN: An adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. *Journal of Artificial Intelligence Research 13*, 155–188.

Cheng, J., R. Greiner, J. Kelly, D. Bell, and W. Liu (2002). Learning bayesian networks from data: An information-theory based approach. *Artificial Intelligence*.

Chesley, G. (1978). Subjective probability elicitation techniques: A performance comparison. *Journal of Accounting Research 16*(2), 225–241.

Chickering, D. (1996a). Learning Bayesian networks is NP-Complete. In D. Fisher and H. Lenz (Eds.), *Learning from Data: Artificial Intelligence and Statistics V*, pp. 121–130. Springer-Verlag.

Chickering, D. (2002a, February). Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research 2*, 445–498.

Chickering, D., D. Geiger, and D. Heckerman (1995, January). Learning Bayesian networks: Search methods and experimental results. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pp. 112–128.

Chickering, D., C. Meek, and D. Heckerman (2003). Large-sample learning of Bayesian networks is hard. In *Proc. 19th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 124–133.

Chickering, D. and J. Pearl (1997). A clinician's tool for analyzing non-compliance. *Computing Science and Statistics 29*, 424–31.

Chickering, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In *Proc. 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 87–98.

Chickering, D. M. (1996b). Learning equivalence classes of Bayesian network structures. In *Proc. 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 150–157.

Chickering, D. M. (2002b, November). Optimal structure identification with greedy search. *Journal of Machine Learning Research 3*, 507–554.

Chickering, D. M. and D. Heckerman (1997). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning 29*, 181–212.

Chickering, D. M., D. Heckerman, and C. Meek (1997). A Bayesian approach to learning Bayesian networks with local structure. In *Proc. 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 80–89.

Chow, C. K. and C. N. Liu (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory 14*, 462–467.

Collins, M. (2002). Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proc. Conference on Empirical Methods in Natural*

Language Processing (EMNLP).

Cooper, G. (1990). Probabilistic inference using belief networks is NP-hard. *Artificial Intelligence 42*, 393–405.

Cooper, G. and E. Herskovits (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning 9*, 309–347.

Cooper, G. and C. Yoo (1999). Causal discovery from a mixture of experimental and observational data. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 116–125.

Cooper, G. F. (1988). A method for using belief networks as influence diagrams. In *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence (UAI)*, pp. 55–63.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein (2001). *Introduction to Algorithms*. Cambridge, Massachusetts: MIT Press. 2nd Edition.

Covaliu, Z. and R. Oliver (1995). Representation and solution of decision problems using sequential decision diagrams. *Management Science 41*(12), 1860–81.

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory*. John Wiley & Sons.

Cowell, R. (2005). Local propagation in conditional gaussian Bayesian networks. *Journal of Machine Learning Research 6*, 1517–1550.

Cowell, R. G., A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. New York: Springer-Verlag.

Cox, R. (2001). *Algebra of Probable Inference*. The Johns Hopkins University Press.

Cozman, F. (2000). Credal networks. *Artificial Intelligence 120*, 199–233.

Csiszàr, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability 3*(1), 146–158.

Culotta, A., M. Wick, R. Hall, and A. McCallum (2007). First-order probabilistic models for coreference resolution. In *Proc. Conference of the North American Association for Computational Linguistics*.

D. Rusakov, D. G. (2005). Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research 6*, 1–35.

Dagum, P. and M. Luby (1993). Appoximating probabilistic inference in Bayesian belief networks in NP-hard. *Artificial Intelligence 60*(1), 141–153.

Dagum, P. and M. Luby (1997). An optimal approximation algorithm for Baysian inference. *Artificial Intelligence 93*(1–2), 1–27.

Daneshkhah, A. (2004). Psychological aspects influencing elicitation of subjective probability. Technical report, University of Sheffield.

Darroch, J. and D. Ratcliff (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics 43*, 1470–1480.

Darwiche, A. (1993). Argument calculus and networks. In *Proc. 9th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 420–27.

Darwiche, A. (2001a). Constant space reasoning in dynamic Bayesian networks. *International Journal of Approximate Reasoning 26*, 161–178.

Darwiche, A. (2001b). Recursive conditioning. *Artificial Intelligence 125*(1–2), 5–41.

Darwiche, A. (2003). A differential approach to inference in Bayesian networks. *Journal of the ACM 50*(3), 280–305.

Darwiche, A. and M. Goldszmidt (1994). On the relation between Kappa calculus and probabilistic reasoning. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Dasgupta, S. (1997). The sample complexity of learning fixed-structure Bayesian networks. *Ma-*

chine Learning 29, 165–180.

Dasgupta, S. (1999). Learning polytrees. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 134–141.

Dawid, A. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B 41*, 1–31.

Dawid, A. (1980). Conditional independence for statistical operations. *Annals of Statistics 8*, 598–617.

Dawid, A. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A 147*(2), 278–292.

Dawid, A. (1992). Applications of a general propagation algorithm for probabilistic expert system. *Statistics and Computing 2*, 25–36.

Dawid, A. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review 70*, 161–189. Corrections p437.

Dawid, A. (2007, September). Fundamentals of statistical causality. Technical Report 279, RSS/EPSRC Graduate Training Programme, University of Sheffield.

Dawid, A., U. Kjærulff, and S. Lauritzen (1995). Hybrid propagation in junction trees. In *Advances in Intelligent Computing*, Volume 945. Springer-Verlag.

de Bombal, F., D. Leaper, J. Staniland, A. McCann, and J. Harrocks (1972). Computer-aided diagnosis of acute abdominal pain. *British Medical Journal 2*, 9–13.

de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. *Annals Institute H. Poincaré 7*, 1–68. Translated by H. Kyburg in Kyburg et al. (1980).

de Freitas, N., P. Højen-Sørensen, M. Jordan, and S. Russell (2001). Variational MCMC. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 120–127.

Dean, T. and K. Kanazawa (1989). A model for reasoning about persistence and causation. *Computational Intelligence 5*(3), 142–150.

Dechter, R. (1997). Mini-Buckets: A general scheme for generating approximations in automated reasoning. In *Proc. 15th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1297–1303.

Dechter, R. (1999). Bucket elimination: A unifying framework for reasoning. *Artificial Intelligence 113*(1–2), 41–85.

Dechter, R. (2003). *Constraint Processing*. Morgan Kaufmann.

Dechter, R., K. Kask, and R. Mateescu (2002). Iterative join-graph propagation. In *Proc. 18th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 128–136.

Dechter, R. and I. Rish (1997). A scheme for approximating probabilistic inference. In *Proc. 13th Conference on Uncertainty in Artificial Intelligence (UAI)*.

DeGroot, M. H. (1989). *Probability and Statistics*. Reading, MA: Addison Wesley.

Della Pietra, S., V. Della Pietra, and J. Lafferty (1997). Inducing features of random fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence 19*(4), 380–393.

Dellaert, F., S. Seitz, C. Thorpe, and S. Thrun (2003). EM, MCMC, and chain flipping for structure from motion with unknown correspondence. *Machine Learning 50*(1–2), 45–71.

Deming, W. and F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics 11*, 427–444.

Dempster, A., N. M. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*(1), 1–22.

Deng, K. and A. Moore (1989). Multiresolution instance-based learning. In *Proc. 14th International*

Joint Conference on Artificial Intelligence (IJCAI), pp. 1233–1239.

Deshpande, A., M. Garofalakis, and M. Jordan (2001). Efficient stepwise selection in decomposable models. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 128–135.

Diez, F. (1993). Parameter adjustment in Bayes networks: The generalized noisy OR-gate. In *Proc. 9th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 99–105.

Dittmer, S. L. and F. V. Jensen (1997). Myopic value of information in influence diagrams. In *Proc. 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 142–149.

Doucet, A. (1998). On sequential simulation-based methods for Bayesian filtering. Technical Report CUED/FINFENG/TR 310, Department of Engineering, Cambridge University.

Doucet, A., N. de Freitas, and N. Gordon (Eds.) (2001). *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

Doucet, A., N. de Freitas, K. Murphy, and S. Russell (2000). Rao-Blackwellised particle filtering for dynamic Bayesian networks. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Doucet, A., S. Godsill, and C. Andrieu (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing 10*(3), 197–208.

Drummond, M., B. O'Brien, G. Stoddart, and G. Torrance (1997). *Methods for the Economic Evaluation of Health Care Programmes, 2nd Edition*. Oxford, UK: Oxford University Press.

Druzdzel, M. (1993). *Probabilistic Reasoning in Decision Support Systems: From Computation to Common Sense*. Ph.D. thesis, Carnegie Mellon University.

Dubois, D. and H. Prade (1990). Inference i possibilistic hypergraphs. In *Proc. of the 6th Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*.

Duchi, J., D. Tarlow, G. Elidan, and D. Koller (2006). Using combinatorial optimization within max-product belief propagation. In *Proc. 20th Conference on Neural Information Processing Systems (NIPS)*.

Duda, R., J. Gaschnig, and P. Hart (1979). Model design in the PROSPECTOR consultant system for mineral exploration. In D. Michie (Ed.), *Expert Systems in the Microelectronic Age*, pp. 153–167. Edinburgh, Scotland: Edinburgh University Press.

Duda, R. and P. Hart (1973). *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons.

Duda, R., P. Hart, and D. Stork (2000). *Pattern Classification, Second Edition*. Wiley.

Dudík, M., S. Phillips, and R. Schapire (2004). Performance guarantees for regularized maximum entropy density estimation. In *Proc. Conference on Computational Learning Theory (COLT)*.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.

Dykstra, R. and J. Lemke (1988). Duality of I projections and maximum likelihood estimation for log-linear models under cone constraints. *Journal of the American Statistical Association 83*(402), 546–554.

El-Hay, T. and N. Friedman (2001). Incorporating expressive graphical models in variational approximations: Chain-graphs and hidden variables. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 136–143.

Elfadel, I. (1995). Convex potentials and their conjugates in analog mean-field optimization. *Neural Computation 7*, 1079–1104.

Elidan, G. and N. Friedman (2005). Learning hidden variable networks: The information bottleneck approach. *Journal of Machine Learning Research 6*, 81–127.

Elidan, G., I. McGraw, and D. Koller (2006). Residual belief propagation: Informed scheduling for asynchronous message passing. In *Proc. 22nd Conf. on Uncertainty in Artificial Intelligence*.

Elidan, G., N. Lotner, N. Friedman, and D. Koller (2000). Discovering hidden variables: A structure-based approach. In *Proc. 14th Conf. on Neural Information Processing Systems (NIPS)*.

Elidan, G., I. Nachman, and N. Friedman (2007). "Ideal Parent" structure learning for continuous variable networks. *Journal of Machine Learning Research 8*, 1799–1833.

Elidan, G., M. Ninio, N. Friedman, and D. Schuurmans (2002). Data perturbation for escaping local maxima in learning. In *Proc. 18th National Conference on Artificial Intelligence (AAAI)*.

Ellis, B. and W. Wong (2008). Learning causal Bayesian network structures from experimental data. *Journal of the American Statistical Association 103*, 778–789.

Elston, R. C. and J. Stewart (1971). A general model for the analysis of pedigree data. *Human Heredity 21*, 523–542.

Ezawa, K. (1994). Value of evidence on influence diagrams. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 212–220.

Feller, W. (1970). *An Introduction to Probability Theory and Its Applications* (third ed.), Volume I. New York: John Wiley & Sons.

Felzenszwalb, P. and D. Huttenlocher (2006, October). Efficient belief propagation for early vision. *International Journal of Computer Vision 70*(1).

Fertig, K. and J. Breese (1989). Interval influence diagrams. In *Proc. 5th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Fine, S., Y. Singer, and N. Tishby (1998). The hierarchical Hidden Markov Model: Analysis and applications. *Machine Learning 32*, 41–62.

Fishburn, P. (1967). Interdependence and additivity in multivariate, unidimensional expected utility theory. *International Economic Review 8*, 335–42.

Fishburn, P. (1970). *Utility Theory for Decision Making*. New York: Wiley.

Fishelson, M. and D. Geiger (2003). Optimizing exact genetic linkage computations. In *Proc. International Conf. on Research in Computational Molecular Biology (RECOMB)*, pp. 114–121.

Fishman, G. (1976, July). Sampling from the gamma distribution on a computer. *Communications of the ACM 19*(7), 407–409.

Fishman, G. (1996). *Monte Carlo — Concept, Algorithms, and Applications*. Series in Operations Research. Springer.

Fox, D., W. Burgard, and S. Thrun (1999). Markov localization for mobile robots in dynamic environments. *Journal of Artificial Intelligence Research 11*, 391–427.

Freund, Y. and R. Schapire (1998). Large margin classification using the perceptron algorithm. In *Proc. Conference on Computational Learning Theory (COLT)*.

Frey, B. (2003). Extending factor graphs so as to unify directed and undirected graphical models. In *Proc. 19th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 257–264.

Frey, B. and A. Kannan (2000). Accumulator networks: suitors of local probability propagation. In *Proc. 14th Conference on Neural Information Processing Systems (NIPS)*.

Frey, B. and D. MacKay (1997). A revolution: Belief propagation in graphs with cycles. In *Proc. 11th Conference on Neural Information Processing Systems (NIPS)*.

Frey, B. J. (1998). *Graphical Models for Machine Learning and Digital Communication*. Cambridge, Massachusetts: MIT Press.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In *Proc. 14th International Conference on Machine Learning (ICML)*, pp. 125–133.

Friedman, N. (1998). The Bayesian structural em algorithm. In *Proc. 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 129–138.

Friedman, N., D. Geiger, and M. Goldszmidt (1997). Bayesian network classifiers. *Machine Learning 29*, 131–163.

Friedman, N., D. Geiger, and N. Lotner (2000). Likelihood computations using value abstraction. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Friedman, N., L. Getoor, D. Koller, and A. Pfeffer (1999). Learning probabilistic relational models. In *Proc. 16th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1300–1307.

Friedman, N. and M. Goldszmidt (1996). Learning Bayesian networks with local structure. In *Proc. 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 252–262.

Friedman, N. and M. Goldszmidt (1998). Learning Bayesian networks with local structure. See Jordan (1998), pp. 421–460.

Friedman, N. and D. Koller (2003). Being Bayesian about Bayesian network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning 50*(1–2), 95–126.

Friedman, N., K. Murphy, and S. Russell (1998). Learning the structure of dynamic probabilistic networks. In *Proc. 14th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Friedman, N. and I. Nachman (2000). Gaussian process networks. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 211–219.

Friedman, N. and Z. Yakhini (1996). On the sample complexity of learning Bayesian networks. In *Proc. 12th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Frogner, C. and A. Pfeffer (2007). Discovering weakly-interacting factors in a complex stochastic process. In *Proc. 21st Conference on Neural Information Processing Systems (NIPS)*.

Frydenberg, J. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics 17*, 790–805.

Fudenberg, D. and J. Tirole (1991). *Game Theory*. MIT Press.

Fung, R. and K. C. Chang (1989). Weighting and integrating evidence for stochastic simulation in Bayesian networks. In *Proc. 5th Conference on Uncertainty in Artificial Intelligence (UAI)*, San Mateo, California. Morgan Kaufmann.

Fung, R. and B. del Favero (1994). Backward simulation in Bayesian networks. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 227–234.

Galles, D. and J. Pearl (1995). Testing identifiability of causal models. In *Proc. 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 185–95.

Gamerman, D. and H. Lopes (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall, CRC.

Ganapathi, V., D. Vickrey, J. Duchi, and D. Koller (2008). Constrained approximate maximum entropy learning. In *Proc. 24th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Garcia, L. D. (2004). Algebraic statistics in model selection. In *Proc. 20th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 177–18.

Geiger, D. and D. Heckerman. A characterization of the bivariate normal-Wishart distribution. *Probability and Mathematical Statistics 18*, 119–131.

Geiger, D. and D. Heckerman (1994). Learning gaussian networks. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 235–243.

Geiger, D. and D. Heckerman (1996). Knowledge representation and inference in similarity networks and Bayesian multinets. *Artificial Intelligence 82*(1-2), 45–74.

Geiger, D., D. Heckerman, H. King, and C. Meek (2001). Stratified exponential families: Graphical

models and model selection. *Annals of Statistics 29*, 505–529.

Geiger, D., D. Heckerman, and C. Meek (1996). Asymptotic model selection for directed networks with hidden variables. In *Proc. 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 283–290.

Geiger, D. and C. Meek (1998). Graphical models and exponential families. In *Proc. 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 156–165.

Geiger, D., C. Meek, and Y. Wexler (2006). A variational inference procedure allowing internal structure for overlapping clusters and deterministic constraints. *Journal of Artificial Intelligence Research 27*, 1–23.

Geiger, D. and J. Pearl (1988). On the logic of causal models. In *Proc. 4th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 3–14.

Geiger, D. and J. Pearl (1993). Logical and algorithmic properties of conditional independence and graphical models. *Annals of Statistics 21*(4), 2001–21.

Geiger, D., T. Verma, and J. Pearl (1989). d-separation: From theorems to algorithms. In *Proc. 5th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 139–148.

Geiger, D., T. Verma, and J. Pearl (1990). Identifying independence in Bayesian networks. *Networks 20*, 507–534.

Gelfand, A. and A. Smith (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 398–409.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Gelman, A. and X.-L. Meng (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science 13*(2), 163–185.

Gelman, A. and D. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*, 457–511.

Geman, S. and D. Geman (1984, November). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence 6*(6), 721–741.

Getoor, L., N. Friedman, D. Koller, A. Pfeffer, and B. Taskar (2007). Probabilistic relational models. See Getoor and Taskar (2007).

Getoor, L., N. Friedman, D. Koller, and B. Taskar (2002). Learning probabilistic models of link structure. *Journal of Machine Learning Research 3*(December), 679–707.

Getoor, L. and B. Taskar (Eds.) (2007). *Introduction to Statistical Relational Learning*. MIT Press.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*, 1317–1339.

Geyer, C. and E. Thompson (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B*.

Geyer, C. and E. Thompson (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association 90*(431), 909–920.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of 23rd Symposium on the Interface Interface Foundation*, pp. 156–163. Fairfax Station.

Ghahramani, Z. (1994). Factorial learning and the em algorithm. In *Proc. 8th Conference on Neural Information Processing Systems (NIPS)*, pp. 617–624.

Ghahramani, Z. and M. Beal (2000). Propagation algorithms for variational Bayesian learning. In

Proc. 14th Conference on Neural Information Processing Systems (NIPS).

Ghahramani, Z. and G. Hinton (1998). Variational learning for switching state-space models. *Neural Computation 12*(4), 963–996.

Ghahramani, Z. and M. Jordan (1993). Supervised learning from incomplete data via an EM approach. In *Proc. 7th Conference on Neural Information Processing Systems (NIPS)*.

Ghahramani, Z. and M. Jordan (1997). Factorial hidden Markov models. *Machine Learning 29*, 245–273.

Gibbs, J. (1902). *Elementary Principles of Statistical Mechanics*. New Haven, Connecticut: Yale University Press.

Gidas, B. (1988). Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbsian distributions. In W. Fleming and P.-L. Lions (Eds.), *Stochastic differential systems, stochastic control theory and applications*. Springer, New York.

Gilks, W. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics 4*, pp. 641–649. Oxford, UK: Clarendon Press.

Gilks, W., N. Best, and K. Tan (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Annals of Statistics 44*, 455–472.

Gilks, W., S. Richardson, and D. Spiegelhalter (Eds.) (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Gilks, W., A. Thomas, and D. Spiegelhalter (1994). A language and program for complex Bayesian modeling. *The Statistician 43*, 169–177.

Gilks, W. and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Annals of Statistics 41*, 337–348.

Giudici, P. and P. Green (1999, December). Decomposable graphical Gaussian model determination. *Biometrika 86*(4), 785–801.

Globerson, A. and T. Jaakkola (2007a). Convergent propagation algorithms via oriented trees. In *Proc. 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*.

Globerson, A. and T. Jaakkola (2007b). Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *Proc. 21st Conference on Neural Information Processing Systems (NIPS)*.

Glover, F. and M. Laguna (1993). Tabu search. In C. Reeves (Ed.), *Modern Heuristic Techniques for Combinatorial Problems*, Oxford, England. Blackwell Scientific Publishing.

Glymour, C. and G. F. Cooper (Eds.) (1999). *Computation, Causation, Discovery*. Cambridge: MIT Press.

Godsill, S., A. Doucet, and M. West (2000). Methodology for Monte Carlo smoothing with application to time-varying autoregressions. In *Proc. International Symposium on Frontiers of Time Series Modelling*.

Golumbic, M. (1980). *Algorithmic Graph Theory and Perfect Graphs*. London: Academic Press.

Good, I. (1950). *Probability and the Weighing of Evidence*. London: Griffin.

Goodman, J. (2004). Exponential priors for maximum entropy models. In *Proc. Conference of the North American Association for Computational Linguistics*.

Goodman, L. (1970). The multivariate analysis of qualitative data: Interaction among multiple classification. *Journal of the American Statistical Association 65*, 226–56.

Gordon, N., D. Salmond, and A. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F 140*(2), 107–113.

Gorry, G. and G. Barnett (1968). Experience with a model of sequential diagnosis. *Computers and Biomedical Research 1*, 490–507.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*, 711–732.

Green, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society, Series B 52*(3), 443–452.

Greig, D., B. Porteous, and A. Seheult (1989). Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society, Series B 51*(2), 271–279.

Greiner, R. and W. Zhou (2002). Structural extension to logistic regression: Discriminant parameter learning of belief net classifiers. In *Proc. 18th Conference on Artificial Intelligence (AAAI)*.

Guestrin, C. E., D. Koller, R. Parr, and S. Venkataraman (2003). Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research 19*, 399–468.

Guyon, X. and H. R. Künsch (1992). Asymptotic comparison of estimators in the Ising model. In *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis, Lecture Notes in Statistics*, Volume 74, pp. 177–198. Springer, Berlin.

Ha, V. and P. Haddawy (1997). Problem-focused incremental elicitation of multi-attribute utility models. In *Proc. 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 215–222.

Ha, V. and P. Haddawy (1999). A hybrid approach to reasoning with partially elicited preference models. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 263–270.

Haberman, S. (1974). *The General Log-Linear Model*. Ph.D. thesis, Department of Statistics, University of Chicago.

Halpern, J. Y. (2003). *Reasoning about Uncertainty*. MIT Press.

Hammer, P. (1965). Some network flow problems solved with pseudo-Boolean programming. *Operations Research 13*, 388–399.

Hammersley, J. and P. Clifford (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.

Handschin, J. and D. Mayne (1969). Monte Carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. *International Journal of Control 9*(5), 547–559.

Hartemink, A., D. Gifford, T. Jaakkola, and R. Young (2002, March/April). Bayesian methods for elucidating genetic regulatory networks. *IEEE Intelligent Systems 17*, 37–43. special issue on Intelligent Systems in Biology.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics.

Hazan, T. and A. Shashua (2008). Convergent message-passing algorithms for inference over general graphs with convex free energies. In *Proc. 24th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Heckerman, D. (1990). *Probabilistic Similarity Networks*. MIT Press.

Heckerman, D. (1993). Causal independence for knowledge acquisition and inference. In *Proc. 9th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 122–127.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. See Jordan (1998).

Heckerman, D. and J. Breese (1996). Causal independence for probability assessment and inference using Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics 26*, 826–831.

Heckerman, D., J. Breese, and K. Rommelse (1995, March). Decision-theoretic troubleshooting.

Communications of the ACM 38(3), 49–57.

Heckerman, D., D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie (2000). Dependency networks for inference, collaborative filtering, and data visualization. *jmlr 1*, 49–75.

Heckerman, D. and D. Geiger (1995). Learning Bayesian networks: a unification for discrete and Gaussian domains. In *Proc. 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 274–284.

Heckerman, D., D. Geiger, and D. M. Chickering (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning 20*, 197–243.

Heckerman, D., E. Horvitz, and B. Nathwani (1992). Toward normative expert systems: Part I. The Pathfinder project. *Methods of Information in Medicine 31*, 90–105.

Heckerman, D. and H. Jimison (1989). A Bayesian perspective on confidence. In *Proc. 5th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 149–160.

Heckerman, D., A. Mamdani, and M. Wellman (1995). Real-world applications of Bayesian networks. *Communications of the ACM 38*.

Heckerman, D. and C. Meek (1997). Embedded Bayesian network classifiers. Technical Report MSR-TR-97-06, Microsoft Research, Redmond, WA.

Heckerman, D., C. Meek, and G. Cooper (1999). A Bayesian approach to causal discovery. See Glymour and Cooper (1999), pp. 141–166.

Heckerman, D., C. Meek, and D. Koller (2007). Probabilistic entity-relationship models, PRMs, and plate models. See Getoor and Taskar (2007).

Heckerman, D. and B. Nathwani (1992a). An evaluation of the diagnostic accuracy of Pathfinder. *Computers and Biomedical Research 25(1)*, 56–74.

Heckerman, D. and B. Nathwani (1992b). Toward normative expert systems. II. Probability-based representations for efficient knowledge acquisition and inference. *Methods of Information in Medicine 31*, 106–16.

Heckerman, D. and R. Shachter (1994). A decision-based view of causality. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 302–310. Morgan Kaufmann.

Henrion, M. (1986). Propagation of uncertainty in Bayesian networks by probabilistic logic sampling. In *Proc. 2nd Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 149–163.

Henrion, M. (1991). Search-based algorithms to bound diagnostic probabilities in very large belief networks. In *Proc. 7th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 142–150.

Hernández, L. and S. Moral (1997). Mixing exact and importance sampling propagation algorithms in dependence graphs. *International Journal of Intelligent Systems 12*, 553–576.

Heskes, T. (2002). Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In *Proc. 16th Conference on Neural Information Processing Systems (NIPS)*, pp. 359–366.

Heskes, T. (2004). On the uniqueness of loopy belief propagation fixed points. *Neural Computation 16*, 2379–2413.

Heskes, T. (2006). Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. *Journal of Machine Learning Research 26*, 153–190.

Heskes, T., K. Albers, and B. Kappen (2003). Approximate inference and constrained optimization. In *Proc. 19th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 313–320.

Heskes, T., M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter (2005). Approximate inference techniques with expectation constraints. *Journal of Statistical Mechanics: Theory and Experiment*.

Heskes, T. and O. Zoeter (2002). Expectation propagation for approximate inference in dynamic

Bayesian networks. In *Proc. 18th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Heskes, T. and O. Zoeter (2003). Generalized belief propagation for approximate inference in hybrid Bayesian networks. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*.

Heskes, T., O. Zoeter, and W. Wiegerinck (2003). Approximate expectation maximization. In *Proc. 17th Conference on Neural Information Processing Systems (NIPS)*, pp. 353–360.

Higdon, D. M. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association 93*, 585–595.

Hinton, G. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation 14*, 1771–1800.

Hinton, G., S. Osindero, and Y. Teh (2006). A fast learning algorithm for deep belief nets. *Neural Computation 18*, 1527–1554.

Hinton, G. and R. Salakhutdinov (2006). Reducing the dimensionality of data with neural networks. *Science 313*, 504–507.

Hinton, G. and T. Sejnowski (1983). Optimal perceptual inference. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 448–453.

Hinton, G. E., P. Dayan, B. Frey, and R. M. Neal (1995). The wake-sleep algorithm for unsupervised neural networks. *Science 268*, 1158–1161.

Höffgen, K. (1993). Learning and robust learning of product distributions. In *Proc. Conference on Computational Learning Theory (COLT)*, pp. 77–83.

Hofmann, R. and V. Tresp (1995). Discovering structure in continuous variables using bayesian networks. In *Proc. 9th Conference on Neural Information Processing Systems (NIPS)*.

Horn, G. and R. McEliece (1997). Belief propagation in loopy bayesian networks: experimental results. In *Proceedings if IEEE International Symposium on Information Theory*, pp. 232.

Horvitz, E. and M. Barry (1995). Display of information for time-critical decision making. In *Proc. 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 296–305.

Horvitz, E., J. Breese, and M. Henrion (1988). Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning 2*, 247–302. Special Issue on Uncertainty in Artificial Intelligence.

Horvitz, E., H. Suermondt, and G. Cooper (1989). Bounded conditioning: Flexible inference for decisions under scarce resources. In *Proc. 5th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 182–193.

Howard, R. (1970). Decision analysis: Perspectives on inference, decision, and experimentation. *Proceedings of the IEEE 58*, 632–643.

Howard, R. (1977). Risk preference. In R. Howard and J. Matheson (Eds.), *Readings in Decision Analysis*, pp. 429–465. Menlo Park, California: Decision Analysis Group, SRI International.

Howard, R. and J. Matheson (1984a). Influence diagrams. See Howard and Matheson (1984b), pp. 721–762.

Howard, R. and J. Matheson (Eds.) (1984b). *The Principle and Applications of Decision Analysis*. Menlo Park, CA, USA: Strategic Decisions Group.

Howard, R. A. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics SSC-2*, 22–26.

Howard, R. A. (1989). Microrisks for medical decision analysis. *International Journal of Technology Assessment in Health Care 5*, 357–370.

Huang, C. and A. Darwiche (1996). Inference in belief networks: A procedural guide. *International*

Journal of Approximate Reasoning 15(3), 225–263.

Huang, F. and Y. Ogata (2002). Generalized pseudo-likelihood estimates for Markov random fields on lattice. *Annals of the Institute of Statistical Mathematics 54*, 1–18.

Ihler, A. (2007). Accuracy bounds for belief propagation. In *Proc. 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*.

Ihler, A. T., J. W. Fisher, and A. S. Willsky (2003). Message errors in belief propagation. In *Proc. 17th Conference on Neural Information Processing Systems (NIPS)*.

Ihler, A. T., J. W. Fisher, and A. S. Willsky (2005). Loopy belief propagation: Convergence and effects of message errors. *Journal of Machine Learning Research 6*, 905–936.

Imoto, S., S. Kim, T. Goto, S. Aburatani, K. Tashiro, S. Kuhara, and S. Miyano (2003). Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *Journal of Bioinformatics and Computational Biology 1*, 231–252.

Indyk, P. (2004). Nearest neighbors in high-dimensional spaces. In J. Goodman and J. O'Rourke (Eds.), *Handbook of Discrete and Computational Geometry* (2nd ed.). CRC Press.

Isard, M. (2003). PAMPAS: Real-valued graphical models for computer vision. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 613–620.

Isard, M. and A. Blake (1998a). Condensation — conditional density propagation for visual tracking. *International Journal of Computer Vision 29*(1), 5–28.

Isard, M. and A. Blake (1998b). A smoothing filter for condensation. In *Proc. European Conference on Computer Vision (ECCV)*, Volume 1, pp. 767–781.

Isham, V. (1981). An introduction to spatial point processes and Markov random fields. *International Statistical Review 49*, 21–43.

Ishikawa, H. (2003). Exact optimization for Markov random fields with convex priors. *IEEE Trans. on Pattern Analysis and Machine Intelligence 25*(10), 1333–1336.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Z. Phys. 31*, 253–258.

Jaakkola, T. (2001). Tutorial on variational approximation methods. In M. Opper and D. Saad (Eds.), *Advanced mean field methods*, pp. 129–160. Cambridge, Massachusetts: MIT Press.

Jaakkola, T. and M. Jordan (1996a). Computing upper and lower bounds on likelihoods in intractable networks. In *Proc. 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 340–348.

Jaakkola, T. and M. Jordan (1996b). Recursive algorithms for approximating probabilities in graphical models. In *Proc. 10th Conference on Neural Information Processing Systems (NIPS)*, pp. 487–93.

Jaakkola, T. and M. Jordan (1997). A variational approach to bayesian logistic regression models and their extensions. In *Proc. 6thWorkshop on Artificial Intelligence and Statistics*.

Jaakkola, T. and M. Jordan (1998). Improving the mean field approximation via the use of mixture models. See Jordan (1998).

Jaakkola, T. and M. Jordan (1999). Variational probabilistic inference and the QMR-DT network. *Journal of Artificial Intelligence Research 10*, 291–322.

Jarzynski, C. (1997, Apr). Nonequilibrium equality for free energy differences. *Physical Review Letters 78*(14), 2690–2693.

Jaynes, E. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jensen, F., F. V. Jensen, and S. L. Dittmer (1994). From influence diagrams to junction trees. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 367–73.

Jensen, F. and M. Vomlelová (2003). Unconstrained influence diagrams. In *Proc. 19th Conference*

on Uncertainty in Artificial Intelligence (UAI), pp. 234–41.

Jensen, F. V. (1995). Cautious propagation in Bayesian networks. In *Proc. 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 323–328.

Jensen, F. V. (1996). *An introduction to Bayesian Networks*. London: University College London Press.

Jensen, F. V., K. G. Olesen, and S. K. Andersen (1990, August). An algebra of Bayesian belief universes for knowledge-based systems. *Networks 20*(5), 637–659.

Jerrum, M. and A. Sinclair (1997). The Markov chain Monte Carlo method. In D. Hochbaum (Ed.), *Approximation Algorithms for NP-hard Problems*. Boston: PWS Publishing.

Ji, C. and L. Seymour (1996). A consistent model selection procedure for Markov random fields based on penalized pseudolikelihood. *Annals of Applied Probability*.

Jimison, H., L. Fagan, R. Shachter, and E. Shortliffe (1992). Patient-specific explanation in models of chronic disease. *AI in Medicine 4*, 191–205.

Jordan, M., Z. Ghahramani, T. Jaakkola, and L. K. Saul (1998). An introduction to variational approximations methods for graphical models. See Jordan (1998).

Jordan, M. I. (Ed.) (1998). *Learning in Graphics Models*. Cambridge, MA: The MIT Press.

Julier, S. (2002). The scaled unscented transformation. In *Proceedings of the American Control Conference*, Volume 6, pp. 4555–4559.

Julier, S. and J. Uhlmann (1997). A new extension of the Kalman filter to nonlinear systems. In *Proc. of AeroSense: The 11th International Symposium on Aerospace/Defence Sensing, Simulation and Controls*.

Kahneman, D., P. Slovic, and A. Tversky (Eds.) (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kalman, R. and R. Bucy (1961). New results in linear filtering and prediction theory. *Trans. ASME, Series D, Journal of Basic Engineering*.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering 82*(Series D), 35–45.

Kanazawa, K., D. Koller, and S. Russell (1995). Stochastic simulation algorithms for dynamic probabilistic networks. In *Proc. 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 346–351.

Kass, R. and A. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Kearns, M., M. L. Littman, and S. Singh (2001). Graphical models for game theory. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 253–260.

Kearns, M. and Y. Mansour (1998). Exact inference of hidden structure from sample data in noisy-or networks. In *Proc. 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 304–31.

Kearns, M., Y. Mansour, and A. Ng (1997). An information-theoretic analysis of hard and soft assignment methods for clustering. In *Proc. 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 282–293.

Keeney, R. L. and H. Raiffa (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley & Sons, Inc.

Kersting, K. and L. De Raedt (2007). Bayesian logic programming: Theory and tool. See Getoor and Taskar (2007).

Kikuchi, R. (1951). A theory of cooperative phenomena. *Physical Review Letters 81*, 988–1003.

Kim, C.-J. and C. Nelson (1998). *State-Space Models with Regime-Switching: Classical and Gibbs-Sampling Approaches with Applications*. MIT Press.

Kim, J. and J. Pearl (1983). A computational model for combined causal and diagnostic reasoning in inference systems. In *Proc. 7th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 190–193.

Kirkpatrick, S., C. Gelatt, and M. Vecchi (1983). Optimization by simulated annealing. *Science 220*, 671–680.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics 5*(1), 1–25.

Kjærulff, U. (1990, March). Triangulation of graph — Algorithms giving small total state space. Technical Report R90-09, Aalborg University, Denmark.

Kjærulff, U. (1992). A computational scheme for reasoning in dynamic probabilistic networks. In *Proc. 8th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 121–129.

Kjærulff, U. (1995a). dHugin: A computational system for dynamic time-sliced Bayesian networks. *International Journal of Forecasting 11*, 89–111.

Kjærulff, U. (1995b). HUGS: Combining exact inference and Gibbs sampling in junction trees. In *Proc. 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 368–375.

Kjaerulff, U. (1997). Nested junction trees. In *Proc. 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 294–301.

Kjærulff, U. and L. van der Gaag (2000). Making sensitivity analysis computationally efficient. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 317–325.

Koivisto, M. and K. Sood (2004). Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research 5*, 549–573.

Kok, J., M. Spaan, and N. Vlassis (2003). Multi-robot decision making using coordination graphs. In *Proc. International Conference on Advanced Robotics (ICAR)*, pp. 1124–1129.

Kok, J. and N. Vlassis (2005). Using the max-plus algorithm for multiagent decision making in coordination graphs. In *RoboCup-2005: Robot Soccer World Cup IX*, Osaka, Japan.

Koller, D. and R. Fratkina (1998). Using learning for approximation in stochastic processes. In *Proc. 15th International Conference on Machine Learning (ICML)*, pp. 287–295.

Koller, D., U. Lerner, and D. Anguelov (1999). A general algorithm for approximate inference and its application to hybrid Bayes nets. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 324–333.

Koller, D. and B. Milch (2001). Multi-agent influence diagrams for representing and solving games. In *Proc. 17th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1027–1034.

Koller, D. and B. Milch (2003). Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior 45*(1), 181–221. Full version of paper in IJCAI '03.

Koller, D. and R. Parr (1999). Computing factored value functions for policies in structured MDPs. In *Proc. 16th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1332–1339.

Koller, D. and A. Pfeffer (1997). Object-oriented Bayesian networks. In *Proc. 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 302–313.

Kolmogorov, V. (2006). Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kolmogorov, V. and C. Rother (2006). Comparison of energy minimization algorithms for highly connected graphs. In *Proc. European Conference on Computer Vision (ECCV)*.

Kolmogorov, V. and M. Wainwright (2005). On the optimality of tree reweighted max-product

message passing. In *Proc. 21st Conference on Uncertainty in Artificial Intelligence (UAI)*.

Kolmogorov, V. and R. Zabih (2004). What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*(2).

Komarek, P. and A. Moore (2000). A dynamic adaptation of AD-trees for efficient machine learning on large data sets. In *Proc. 17th International Conference on Machine Learning (ICML)*, pp. 495–502.

Komodakis, N., N. Paragios, and G. Tziritas (2007). MRF optimization via dual decomposition: Message-passing revisited. In *Proc. International Conference on Computer Vision (ICCV)*.

Komodakis, N. and G. Tziritas (2005). A new framework for approximate labeling via graph-cuts. In *Proc. International Conference on Computer Vision (ICCV)*.

Komodakis, N., G. Tziritas, and N. Paragios (2007). Fast, approximately optimal solutions for single and dynamic MRFs. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kong, A. (1991). Efficient methods for computing linkage likelihoods of recessive diseases in inbred pedigrees. *Genetic Epidemiology 8*, 81–103.

Korb, K. and A. Nicholson (2003). *Bayesian Artificial Intelligence*. CRC Press.

Koster, J. (1996). Markov properties of non-recursive causal models. *The Annals of Statistics 24*(5), 2148–77.

Kočka, T., R. Bouckaert, and M. Studený (2001). On characterizing inclusion of Bayesian networks. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 261–68.

Kozlov, A. and D. Koller (1997). Nonuniform dynamic discretization in hybrid networks. In *Proc. 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 314–325.

Krause, A. and C. Guestrin (2005a). Near-optimal nonmyopic value of information in graphical models. In *Proc. 21st Conference on Uncertainty in Artificial Intelligence (UAI)*.

Krause, A. and C. Guestrin (2005b). Optimal nonmyopic value of information in graphical models: Efficient algorithms and theoretical limits. In *Proc. 19th International Joint Conference on Artificial Intelligence (IJCAI)*.

Kreps, D. (1988). *Notes on the Theory of Choice*. Boulder, Colorado: Westview Press.

Kschischang, F. and B. Frey (1998). Iterative decoding of compound codes by probability propagation in graphical models. *IEEE Journal on Selected Areas in Communications 16*, 219–230.

Kschischang, F., B. Frey, and H.-A. Loeliger (2001a). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory 47*, 498–519.

Kschischang, F., B. Frey, and H.-A. Loeliger (2001b). Factor graphs and the sum-product algorithm. *IEEE Trans. Information Theory 47*, 498–519.

Kullback, S. (1959). *Information Theory and Statistics*. New York: John Wiley & Sons.

Kumar, M., V. Kolmogorov, and P. Torr (2007). An analysis of convex relaxations for MAP estimation. In *Proc. 21st Conference on Neural Information Processing Systems (NIPS)*.

Kumar, M., P. Torr, and A. Zisserman (2006). Solving Markov random fields using second order cone programming relaxations. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1045–1052.

Kuppermann, M., S. Shiboski, D. Feeny, E. Elkin, and A. Washington (1997, Jan–Mar). Can preference scores for discrete states be used to derive preference scores for an entire path of events? An application to prenatal diagnosis. *Medical Decision Making 17*(1), 42–55.

Kyburg, H., , and H. Smokler (Eds.) (1980). *Studies in Subjective Probability*. New York: Krieger.

La Mura, P. (2000). Game networks. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence*

(UAI), pp. 335–342.

La Mura, P. and Y. Shoham (1999). Expected utility networks. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 366–73.

Lacoste-Julien, S., B. Taskar, D. Klein, and M. Jordan (2006, June). Word alignment via quadratic assignment. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pp. 112–119.

Lafferty, J., A. McCallum, and F. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conference on Machine Learning (ICML)*.

Lam, W. and F. Bacchus (1993). Using causal information and local measures to learn Bayesian networks. In *Proc. 9th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 243–250.

Lange, K. and R. C. Elston (1975). Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. *Human Heredity 25*, 95–105.

Laskey, K. (1995). Sensitivity analysis for probability assessments in Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics 25*(6), 901 – 909.

Lauritzen, S. (1982). *Lectures on contingency tables* (2 ed.). Aalborg: Denmark: University of Aalborg Press.

Lauritzen, S. (1992). Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association 87*(420), 1089–1108.

Lauritzen, S. (1996). *Graphical Models*. New York: Oxford University Press.

Lauritzen, S. and D. Nilsson (2001). Representing and solving decision problems with limited information. *Management Science 47*(9), 1235–51.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis 19*, 191–201.

Lauritzen, S. L. and F. Jensen (2001). Stable local computation with conditional Gaussian distributions. *Statistics and Computing 11*, 191–203.

Lauritzen, S. L. and D. J. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B B 50*(2), 157–224.

Lauritzen, S. L. and N. Wermuth (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics 17*, 31–57.

LeCun, Y., S. Chopra, R. Hadsell, R. Marc'Aurelio, and F.-J. Huang (2007). A tutorial on energy-based learning. In G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and S. Vishwanathan (Eds.), *Predicting Structured Data*. MIT Press.

Lee, S.-I., V. Ganapathi, and D. Koller (2006). Efficient structure learning of Markov networks using L1-regularization. In *Proc. 20th Conference on Neural Information Processing Systems (NIPS)*.

Lehmann, E. and J. Romano (2008). *Testing Statistical Hypotheses*. Springer Texts in Statistics.

Leisink, M. A. R. and H. J. Kappen (2003). Bound propagation. *Journal of Artificial Intelligence Research 19*, 139–154.

Lerner, U. (2002). *Hybrid Bayesian Networks for Reasoning about Complex Systems*. Ph.D. thesis, Stanford University.

Lerner, U., B. Moses, M. Scott, S. McIlraith, and D. Koller (2002). Monitoring a complex physical system using a hybrid dynamic Bayes net. In *Proc. 18th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 301–310.

Lerner, U. and R. Parr (2001). Inference in hybrid networks: Theoretical limits and practical algorithms. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 310–318.

Lerner, U., R. Parr, D. Koller, and G. Biswas (2000). Bayesian fault detection and diagnosis in dynamic systems. In *Proc. 16th Conference on Artificial Intelligence (AAAI)*, pp. 531–537.

Lerner, U., E. Segal, and D. Koller (2001). Exact inference in networks with discrete children of continuous parents. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 319–328.

Li, S. (2001). *Markov Random Field Modeling in Image Analysis*. Springer.

Liang, P. and M. Jordan (2008). An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *Proc. 25th International Conference on Machine Learning (ICML)*.

Little, R. J. A. (1976). Inference about means for incomplete multivariate data. *Biometrika 63*, 593–604.

Little, R. J. A. and D. B. Rubin (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Liu, D. and J. Nocedal (1989). On the limited memory method for large scale optimization. *Mathematical Programming 45*(3), 503–528.

Liu, J., W. Wong, and A. Kong (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and sampling schemes. *Biometrika 81*, 27–40.

Loomes, G. and R. Sugden (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal 92*, 805–824.

MacEachern, S. and L. Berliner (1994, August). Subsampling the Gibbs sampler. *The American Statistician 48*(3), 188–190.

MacKay, D. J. C. (1997). Ensemble learning for hidden markov models. Unpublished manuscripts, http://wol.ra.phy.cam.ac.uk/mackay.

MacKay, D. J. C. and R. M. Neal (1996). Near shannon limit performance of low density parity check codes. *Electronics Letters 32*, 1645–1646.

Madigan, D., S. Andersson, M. Perlman, and C. Volinsky (1996). Bayesian model averaging and model selection for Markov equivalence classes of acyclic graphs. *Communications in Statistics: Theory and Methods 25*, 2493–2519.

Madigan, D. and E. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association 89*, 1535–1546.

Madigan, D. and J. York (1995). Bayesian graphical models for discrete data. *International statistical Review 63*, 215–232.

Madsen, A. and D. Nilsson (2001). Solving influence diagrams using HUGIN, Shafer-Shenoy and lazy propagation. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 337–45.

Malioutov, D., J. Johnson, and A. Willsky (2006). Walk-sums and belief propagation in Gaussian graphical models. *Journal of Machine Learning Research 7*, 2031–64.

Maneva, E., E. Mossel, and M. Wainwright (2007, July). A new look at survey propagation and its generalizations. *Journal of the ACM 54*(4), 2–41.

Manning, C. and H. Schuetze (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Marinari, E. and G. Parisi (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters 19*, 451.

Marinescu, R., K. Kask, and R. Dechter (2003). Systematic vs. non-systematic algorithms for solving the MPE task. In *Proc. 19th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Marthi, B., H. Pasula, S. Russell, and Y. Peres (2002). Decayed MCMC filtering. In *Proc. 18th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Martin, J. and K. VanLehn (1995). Discrete factor analysis: Learning hidden variables in Bayesian networks. Technical report, Department of Computer Science, University of Pittsburgh.

McCallum, A. (2003). Efficiently inducing features of conditional random fields. In *Proc. 19th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 403–10.

McCallum, A., C. Pal, G. Druck, and X. Wang (2006). Multi-conditional learning: Generative/discriminative training for clustering and classification. In *Proc. 22nd Conference on Artificial Intelligence (AAAI)*.

McCallum, A. and B. Wellner (2005). Conditional models of identity uncertainty with application to noun coreference. In *Proc. 19th Conference on Neural Information Processing Systems (NIPS)*, pp. 905–912.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. London: Chapman & Hall.

McEliece, R., D. MacKay, and J.-F. Cheng (1998, February). Turbo decoding as an instance of Pearl's "belief propagation" algorithm. *IEEE Journal on Selected Areas in Communications 16*(2).

McEliece, R. J., E. R. Rodemich, and J.-F. Cheng (1995). The turbo decision algorithm. In *Proc. 33rd Allerton Conference on Communication Control and Computing*, pp. 366–379.

McLachlan, G. J. and T. Krishnan (1997). *The EM Algorithm and Extensions*. Wiley Interscience.

Meek, C. (1995a). Causal inference and causal explanation with background knowledge. In *Proc. 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 403–418.

Meek, C. (1995b). Strong completeness and faithfulness in Bayesian networks. In *Proc. 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 411–418.

Meek, C. (1997). *Graphical Models: Selecting causal and statistical models*. Ph.D. thesis, Carnegie Mellon University.

Meek, C. (2001). Finding a path is harder than finding a tree. *Journal of Artificial Intelligence Research 15*, 383–389.

Meek, C. and D. Heckerman (1997). Structure and parameter learning for causal independence and causal interaction models. In *Proc. 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 366–375.

Meila, M. and T. Jaakkola (2000). Tractable Bayesian learning of tree belief networks. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Meila, M. and M. Jordan (2000). Learning with mixtures of trees. *Journal of Machine Learning Research 1*, 1–48.

Meltzer, T., C. Yanover, and Y. Weiss (2005). Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 428–435.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equation of state calculation by fast computing machines. *Journal of Chemical Physics 21*, 1087–1092.

Meyer, J., M. Phillips, P. Cho, I. Kalet, and J. Doctor (2004). Application of influence diagrams to prostate intensity-modulated radiation therapy plan selection. *Physics in Medicine and Biology 49*, 1637–53.

Middleton, B., M. Shwe, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper (1991). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base.

II. Evaluation of diagnostic performance. *Methods of Information in Medicine 30*, 256–67.

Milch, B., B. Marthi, and S. Russell (2004). BLOG: Relational modeling with unknown objects. In *ICML 2004 Workshop on Statistical Relational Learning and Its Connections to Other Fields*.

Milch, B., B. Marthi, S. Russell, D. Sontag, D. Ong, and A. Kolobov (2005). BLOG: Probabilistic models with unknown objects. In *Proc. 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1352–1359.

Milch, B., B. Marthi, S. Russell, D. Sontag, D. Ong, and A. Kolobov (2007). BLOG: Probabilistic models with unknown objects. See Getoor and Taskar (2007).

Miller, R., H. Pople, and J. Myers (1982). Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine 307*, 468–76.

Minka, T. (2005). Discriminative models, not discriminative training. Technical Report MSR-TR-2005-144, Microsoft Research.

Minka, T. and J. Lafferty (2002). Expectation propagation for the generative aspect model. In *Proc. 18th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Minka, T. P. (2001a). Algorithms for maximum-likelihood logistic regression. Available from `http://www.stat.cmu.edu/~minka/papers/logreg.html`.

Minka, T. P. (2001b). Expectation propagation for approximate Bayesian inference. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 362–369.

Møller, J. M., A. Pettitt, K. Berthelsen, and R. Reeves (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalisation constants. *Biometrika 93*(2), 451–458.

Montemerlo, M., S. Thrun, D. Koller, and B. Wegbreit (2002). FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *Proc. 18th Conference on Artificial Intelligence (AAAI)*, pp. 593–598.

Monti, S. and G. F. Cooper (1997). Learning Bayesian belief networks with neural network estimators. In *Proc. 11th Conference on Neural Information Processing Systems (NIPS)*, pp. 579–584.

Mooij, J. M. and H. J. Kappen (2007). Sufficient conditions for convergence of the sum-product algorithm. *IEEE Trans. Information Theory 53*, 4422–4437.

Moore, A. (2000). The anchors hierarchy: Using the triangle inequality to survive high-dimensional data. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 397–405.

Moore, A. and W.-K. Wong (2003). Optimal reinsertion: A new search operator for accelerated and more accurate bayesian network structure learning. In *Proc. 20th International Conference on Machine Learning (ICML)*, pp. 552–559.

Moore, A. W. and M. S. Lee (1997). Cached sufficient statistics for efficient machine learning with large datasets. *Journal of Artificial Intelligence Research 8*, 67–91.

Morgan, M. and M. Henrion (Eds.) (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press.

Motwani, R. and P. Raghavan (1995). *Randomized Algorithnms*. Cambridge University Press.

Muramatsu, M. and T. Suzuki (2003). A new second-order cone programming relaxation for max-cut problems. *Journal of Operations Research of Japan 43*, 164–177.

Murphy, K. (1999). Bayesian map learning in dynamic environments. In *Proc. 13th Conference on Neural Information Processing Systems (NIPS)*.

Murphy, K. (2002). Dynamic Bayesian Networks: A tutorial. Technical report, Mas-

sachussetts Institute of Technology. Available from http://www.cs.ubc.ca/~murphyk/Papers/dbnchapter.pdf.

Murphy, K. and M. Paskin (2001). Linear time inference in hierarchical HMMs. In *Proc. 15th Conference on Neural Information Processing Systems (NIPS)*.

Murphy, K. and Y. Weiss (2001). The factored frontier algorithm for approximate inference in DBNs. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Murphy, K. P. (1998). Inference and learning in hybrid Bayesian networks. Technical Report UCB/CSD-98-990, University of California, Berkeley.

Murphy, K. P., Y. Weiss, and M. Jordan (1999). Loopy belief propagation for approximate inference: an empirical study. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 467–475.

Murray, I. and Z. Ghahramani (2004). Bayesian learning in undirected graphical models: Approximate MCMC algorithms. In *Proc. 20th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Murray, I., Z. Ghahramani, and D. MacKay (2006). MCMC for doubly-intractable distributions. In *Proc. 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*.

Myers, J., K. Laskey, and T. Levitt (1999). Learning Bayesian networks from incomplete data with stochastic search algorithms. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 476–485.

Narasimhan, M. and J. Bilmes (2004). PAC-learning bounded tree-width graphical models. In *Proc. 20th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Ndilikilikesha, P. (1994). Potential influence diagrams. *International Journal of Approximate Reasoning 10*, 251–85.

Neal, R. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing 6*, 353–366.

Neal, R. (2001). Annealed importance sampling. *Statistics and Computing 11*(2), 25–139.

Neal, R. (2003). Slice sampling. *Annals of Statistics 31*(3), 705–767.

Neal, R. M. (1992). Asymmetric parallel Boltzmann machines are belief networks. *Neural Computation 4*(6), 832–834.

Neal, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto.

Neal, R. M. and G. E. Hinton (1998). A new view of the EM algorithm that justifies incremental and other variants. See Jordan (1998).

Neapolitan, R. E. (2003). *Learning Bayesian Networks*. Prentice Hall.

Ng, A. and M. Jordan (2000). Approximate inference algorithms for two-layer Bayesian networks. In *Proc. 14th Conference on Neural Information Processing Systems (NIPS)*.

Ng, A. and M. Jordan (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Proc. 16th Conference on Neural Information Processing Systems (NIPS)*.

Ng, B., L. Peshkin, and A. Pfeffer (2002). Factored particles for scalable monitoring. In *Proc. 18th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 370–377.

Ngo, L. and P. Haddawy (1996). Answering queries from context-sensitive probabilistic knowledge bases. *Theoretical Computer Science*.

Nielsen, J., T. Kočka, and J. M. Peña (2003). On local optima in learning Bayesian networks. In *Proc. 19th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 435–442.

Nielsen, T. and F. Jensen (1999). Welldefined decision scenarios. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 502–11.

Nielsen, T. and F. Jensen (2000). Representing and solving asymmetric Bayesian decision problems. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 416–25.

Nielsen, T., P.-H. Wuillemin, F. Jensen, and U. Kjærulff (2000). Using robdds for inference in Bayesian networks with troubleshooting as an example. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 426–35.

Nilsson, D. (1998). An efficient algorithm for finding the M most probable configurations in probabilistic expert systems. *Statistics and Computing 8*(2), 159–173.

Nilsson, D. and S. Lauritzen (2000). Evaluating influence diagrams with LIMIDs. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 436–445.

Nodelman, U., C. R. Shelton, and D. Koller (2002). Continuous time Bayesian networks. In *Proc. 18th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 378–387.

Nodelman, U., C. R. Shelton, and D. Koller (2003). Learning continuous time Bayesian networks. In *Proc. 19th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Norman, J., Y. Shahar, M. Kuppermann, and B. Gold (1998). Decision-theoretic analysis of prenatal testing strategies. Technical Report SMI-98-0711, Stanford University, Section on Medical Informatics.

Normand, S.-L. and D. Tritchler (1992). Parameter updating in a Bayes network. *Journal of the American Statistical Association 87*, 1109–1115.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press.

Nummelin, E. (2002). Mc's for mcmc'ists. *International Statistical Review 70*(2), 215–240.

Olesen, K. G., U. Kjærulff, F. Jensen, B. Falck, S. Andreassen, and S. Andersen (1989). A Munin network for the median nerve — A case study on loops. *Applied Artificial Intelligence 3*, 384–403.

Oliver, R. M. and J. Q. Smith (Eds.) (1990). *Influence Diagrams, Belief Nets and Decision Analysis*. New York: John Wiley & Sons.

Olmsted, S. (1983). *On Representing and Solving Influence Diagrams*. Ph.D. thesis, Stanford University.

Opper, M. and O. Winther (2005). Expectation consistent free energies for approximate inference. In *Proc. 19th Conference on Neural Information Processing Systems (NIPS)*.

Ortiz, L. and L. Kaelbling (1999). Accelerating em: An empirical study. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 512–521.

Ortiz, L. E. and L. P. Kaelbling (2000). Adaptive importance sampling for estimation in structured domains. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 446–454.

Osborne, M. and A. Rubinstein (1994). *A Course in Game Theory*. The MIT Press.

Ostendorf, M., V. Digalakis, and O. Kimball (1996). From HMMs to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing 4*(5), 360–378.

Pakzad, P. and V. Anantharam (2002). Minimal graphical representation of Kikuchi regions. In *Proc. 40th Allerton Conference on Communication Control and Computing*, pp. 1585–1594.

Papadimitriou, C. (1993). *Computational Complexity*. Addison Wesley.

Parisi, G. (1988). *Statistical Field Theory*. Reading, Massachusetts: Addison-Wesley.

Park, J. (2002). MAP complexity results and approximation methods. In *Proc. 18th Conference on*

Uncertainty in Artificial Intelligence (UAI), pp. 388–396.

Park, J. and A. Darwiche (2001). Approximating MAP using local search. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 403–âĂŞ410.

Park, J. and A. Darwiche (2003). Solving MAP exactly using systematic search. In *Proc. 19th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Park, J. and A. Darwiche (2004a). Complexity results and approximation strategies for MAP explanations. *Journal of Artificial Intelligence Research 21*, 101–133.

Park, J. and A. Darwiche (2004b). A differential semantics for jointree algorithms. *Artificial Intelligence 156*, 197–216.

Parter, S. (1961). The user of linear graphs in Gauss elimination. *SIAM Review 3*, 119–130.

Paskin, M. (2003a). Sample propagation. In *Proc. 17th Conference on Neural Information Processing Systems (NIPS)*.

Paskin, M. (2003b). Thin junction tree filters for simultaneous localization and mapping. In *Proc. 18th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1157–1164.

Pasula, H., B. Marthi, B. Milch, S. Russell, and I. Shpitser (2002). Identity uncertainty and citation matching. In *Proc. 16th Conference on Neural Information Processing Systems (NIPS)*, pp. 1401–1408.

Pasula, H., S. Russell, M. Ostland, and Y. Ritov (1999). Tracking many objects with many sensors. In *Proc. 16th International Joint Conference on Artificial Intelligence (IJCAI)*.

Patrick, D., J. Bush, and M. Chen (1973). Methods for measuring levels of well-being for a health status index. *Health Services Research 8*, 228–45.

Pearl, J. (1986a). A constraint-propagation approach to probabilistic reasoning. In *Proc. 2nd Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 357–370.

Pearl, J. (1986b). Fusion, propagation and structuring in belief networks. *Artificial Intelligence 29*(3), 241–88.

Pearl, J. (1987). Evidential reasoning using stochastic simulation of causal models. *Artificial Intelligence 32*, 245–257.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, California: Morgan Kaufmann.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*, 669–710.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press.

Pearl, J. and R. Dechter (1996). Identifying independencies in causal graphs with feedback. In *Proc. 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 420–26.

Pearl, J. and A. Paz (1987). GRAPHOIDS: A graph-based logic for reasoning about relevance relations. In B. Du Boulay, D. Hogg, and L. Steels (Eds.), *Advances in Artificial Intelligence*, Volume 2, pp. 357–363. Amsterdam: North Holland.

Pearl, J. and T. S. Verma (1991). A theory of inferred causation. In *Proc. Conference on Knowledge Representation and Reasoning (KR)*, pp. 441–452.

Pe'er, D., A. Regev, G. Elidan, and N. Friedman (2001). Inferring subnetworks from preturbed expression profiles. *Bioinformatics 17*, S215–S224.

Peng, Y. and J. Reggia (1986). Plausibility of diagnostic hypotheses. In *Proc. 2nd Conference on Artificial Intelligence (AAAI)*, pp. 140–45.

Perkins, S., K. Lacker, and J. Theiler (2003, March). Grafting: Fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learning Research 3*, 1333–1356.

Peterson, C. and J. R. Anderson (1987). A mean field theory learning algorithm for neural

networks. *Complex Systems 1*, 995–1019.

Pfeffer, A., D. Koller, B. Milch, and K. Takusagawa (1999). SPOOK: A system for probabilistic object-oriented knowledge representation. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 541–550.

Poh, K. and E. Horvitz (2003). Reasoning about the value of decision-model refinement: Methods and application. In *Proc. 19th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 174–182.

Poland, W. (1994). *Decision Analysis with Continuous and Discrete Variables: A Mixture Distribution Approach*. Ph.D. thesis, Department of Engineering-Economic Systems, Stanford University.

Poole, D. (1989). Average-case analysis of a search algorithm for estimating prior and posterior probabilities in Bayesian networks with extreme probabilities. In *Proc. 13th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 606–612.

Poole, D. (1993a). Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence 64*(1), 81–129.

Poole, D. (1993b). The use of conflicts in searching Bayesian networks. In *Proc. 9th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 359–367.

Poole, D. and N. Zhang (2003). Exploiting causal independence in Bayesian network inference. *Journal of Artificial Intelligence Research 18*, 263–313.

Poon, H. and P. Domingos (2007). Joint inference in information extraction. In *Proc. 23rd Conference on Artificial Intelligence (AAAI)*, pp. 913–918.

Pradhan, M. and P. Dagum (1996). Optimal Monte Carlo estimation of belief network inference. In *Proc. 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 446–453.

Pradhan, M., M. Henrion, G. Provan, B. Del Favero, and K. Huang (1996). The sensitivity of belief networks to imprecise probabilities: An experimental investigation. *Artificial Intelligence 85*, 363–97.

Pradhan, M., G. M. Provan, B. Middleton, and M. Henrion (1994). Knowledge engineering for large belief networks. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 484–490.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, New York.

Qi, R., N. Zhang, and D. Poole (1994). Solving asymmetric decision problems with influence diagrams. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 491–497.

Qi, Y., M. Szummer, and T. Minka (2005). Bayesian conditional random fields. In *Proc. 11th Workshop on Artificial Intelligence and Statistics*.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286.

Rabiner, L. R. and B. H. Juang (1986, January). An introduction to hidden Markov models. *IEEE ASSP Magazine*, 4–15.

Ramsey, F. (1931). *The Foundations of Mathematics and other Logical Essays*. London: Kegan, Paul, Trench, Trubner & Co., New York: Harcourt, Brace and Company. edited by R.B. Braithwaite.

Rasmussen, C. and C. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Rasmussen, C. E. (1999). The infinite gaussian mixture model. In *Proc. 13th Conference on Neural Information Processing Systems (NIPS)*, pp. 554–560.

Ravikumar, P. and J. Lafferty (2006). Quadratic programming relaxations for metric labelling and Markov random field MAP estimation. In *Proc. 23rd International Conference on Machine*

Learning (ICML).

Renooij, S. and L. van der Gaag (2002). From qualitative to quantitative probabilistic networks. In *Proc. 18th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 422–429.

Richardson, M. and P. Domingos (2006). Markov logic networks. *Machine Learning 62*, 107–136.

Richardson, T. (1994). Properties of cyclic graphical models. Master's thesis, Carnegie Mellon University.

Riezler, S. and A. Vasserman (2004). Incremental feature selection and l1 regularization for relaxed maximum-entropy modeling. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ripley, B. D. (1987). *Stochastic Simulation*. New York: John Wiley & Sons.

Rissanen, J. (1987). Stochastic complexity (with discussion). *Journal of the Royal Statistical Society, Series B 49*, 223–265.

Ristic, B., S. Arulampalam, and N. Gordon (2004). *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House Publishers.

Robert, C. and G. Casella (1996). Rao-Blackwellisation of sampling schemes. *Biometrika 83*(1), 81–94.

Robert, C. and G. Casella (2005). *Monte Carlo Statistical Methods* (2nd ed.). Springer Texts in Statistics.

Robins, J. M. and L. A. Wasserman (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proc. 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 409–420.

Rose, D. (1970). Triangulated graphs and the elimination process. *Journal of Mathematical Analysis and Applications 32*, 597–609.

Ross, S. M. (1988). *A First Course in Probability* (third ed.). London: Macmillan.

Rother, C., S. Kumar, V. Kolmogorov, and A. Blake (2005). Digital tapestry. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Rubin, D. R. (1976). Inference and missing data. *Biometrika 63*, 581–592.

Rusmevichientong, P. and B. Van Roy (2001). An analysis of belief propagation on the turbo decoding graph with Gaussian densities. *IEEE Transactions on Information Theory 48*(2).

Russell, S. and P. Norvig (2003). *Artificial Intelligence: A Modern Approach* (2 ed.). Prentice Hall.

Rustagi, J. (1976). *Variational Methods in Statistics*. New York: Academic Press.

Sachs, K., O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan (2005, April). Causal protein-signaling networks derived from multiparameter single-cell data. *Science 308*(5721), 523–529.

Sakurai, J. J. (1985). *Modern Quantum Mechanics*. Reading, Massachusetts: Addison-Wesley.

Santos, A. (1994). A linear constraint satisfaction approach to cost-based abduction. *Artificial Intelligence 65*(1), 1–28.

Santos, E. (1991). On the generation of alternative explanations with implications for belief revision. In *Proc. 7th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 339–347.

Saul, L., T. Jaakkola, and M. Jordan (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research 4*, 61–76.

Saul, L. and M. Jordan (1999). Mixed memory Markov models: Decomposing complex stochastic processes as mixture of simpler ones. *Machine Learning 37*(1), 75–87.

Saul, L. K. and M. I. Jordan (1996). Exploiting tractable substructures in intractable networks. In

Proc. 10th Conference on Neural Information Processing Systems (NIPS).

Savage, L. (1951). The theory of statistical decision. *Journal of the American Statistical Association 46*, 55–67.

Savage, L. J. (1954). *Foundations of Statistics*. New York: John Wiley & Sons.

Schäffer, A. (1996). Faster linkage analysis computations for pedigrees with loops or unused alleles. *Human Heredity*, 226–235.

Scharstein, D. and R. Szeliski (2003). High-accuracy stereo depth maps using structured light. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, Volume 1, pp. 195–202.

Schervish, M. (1995). *Theory of Statistics*. Springer-Verlag.

Schlesinger, M. (1976). Sintaksicheskiy analiz dvumernykh zritelnikh singnalov v usloviyakh pomekh (syntactic analysis of two-dimensional visual signals in noisy conditions). *Kibernetika 4*, 113–130.

Schlesinger, M. and V. Giginyak (2007a). Solution to structural recognition (max,+)-problems by their equivalent transformations (part 1). *Control Systems and Computers 1*, 3–15.

Schlesinger, M. and V. Giginyak (2007b). Solution to structural recognition (max,+)-problems by their equivalent transformations (part 2). *Control Systems and Computers 2*, 3–18.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Segal, E., D. Pe'er, A. Regev, D. Koller, and N. Friedman (2005, April). Learning module networks. *Journal of Machine Learning Research 6*, 557–588.

Segal, E., B. Taskar, A. Gasch, N. Friedman, and D. Koller (2001). Rich probabilistic models for gene expression. *Bioinformatics 17*(Suppl 1), S243–52.

Settimi, R. and J. Smith (2000). Geometry, moments and conditional independence trees with hidden variables. *Annals of Statistics*.

Settimi, R. and J. Q. Smith (1998a). On the geometry of Bayesian graphical models with hidden variables. In *Proc. 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 472–479.

Settimi, R. and J. Q. Smith (1998b). On the geometry of Bayesian graphical models with hidden variables. In *Proc. 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 472–479.

Shachter, R. (1988, July–August). Probabilistic inference and influence diagrams. *Operations Research 36*, 589–605.

Shachter, R. (1999). Efficient value of information computation. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 594–601.

Shachter, R., S. K. Andersen, and P. Szolovits (1994). Global conditioning for probabilistic inference in belief networks. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 514–522.

Shachter, R. and D. Heckerman (1987). Thinking backwards for knowledge acquisition. *Artificial Intelligence Magazine 8*, 55 – 61.

Shachter, R. and C. Kenley (1989). Gaussian influence diagrams. *Management Science 35*, 527–550.

Shachter, R. and P. Ndilikilikesha (1993). Using influence diagrams for probabilistic inference and decision making. In *Proc. 9th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 276–83.

Shachter, R. D. (1986). Evaluating influence diagrams. *Operations Research 34*, 871–882.

Shachter, R. D. (1989). Evidence absorption and propagation through evidence reversals. In *Proc. 5th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 173–190.

Shachter, R. D. (1998). Bayes-ball: The rational pastime. In *Proc. 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 480–487.

Shachter, R. D., B. D'Ambrosio, and B. A. Del Favero (1990). Symbolic probabilistic inference in belief networks. In *Proc. 6th Conference on Artificial Intelligence (AAAI)*, pp. 126–131.

Shachter, R. D. and M. A. Peot (1989). Simulation approaches to general probabilistic inference on belief networks. In *Proc. 5th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 221–230.

Shachter, R. D. and M. A. Peot (1992). Decision making using probabilistic inference methods. In *Proc. 8th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 276–83.

Shafer, G. and J. Pearl (Eds.) (1990). *Readings in Uncertain Reasoning*. Representation and Reasoning. San Mateo, California: Morgan Kaufmann.

Shafer, G. and P. Shenoy (1990). Probability propagation. *Annals of Mathematics and Artificial Intelligence 2*, 327–352.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal 27*, 379–423; 623–656.

Shawe-Taylor, J. and N. Cristianini (2000). *Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

Shenoy, P. (1989). A valuation-based language for expert systems. *International Journal of Approximate Reasoning 3*, 383–411.

Shenoy, P. (2000). Valuation network representation and solution of asymmetric decision problems. *European Journal of Operational Research 121*(3), 579–608.

Shenoy, P. and G. Shafer (1990). Axioms for probability and belief-function propagation. In *Proc. 6th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 169–198.

Shenoy, P. P. (1992). Valuation-based systems for Bayesian decision analysis. *Operations Research 40*, 463–484.

Shental, N., A. Zomet, T. Hertz, and Y. Weiss (2003). Learning and inferring image segmentations using the GBP typical cut algorithm. In *Proc. International Conference on Computer Vision*.

Shimony, S. (1991). Explanation, irrelevance and statistical independence. In *Proc. 7th Conference on Artificial Intelligence (AAAI)*.

Shimony, S. (1994). Finding MAPs for belief networks in NP-hard. *Artificial Intelligence 68*(2), 399–410.

Shoikhet, K. and D. Geiger (1997). A practical algorithm for finding optimal triangulations. In *Proc. 13th Conference on Artificial Intelligence (AAAI)*, pp. 185–190.

Shwe, M. and G. Cooper (1991). An empirical analysis of likelihood-weighting simulation on a large, multiply connected medical belief network. *Computers and Biomedical Research 24*, 453–475.

Shwe, M., B. Middleton, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper (1991). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I. The probabilistic model and inference algorithms. *Methods of Information in Medicine 30*, 241–55.

Silander, T. and P. Myllymaki (2006). A simple approach for finding the globally optimal Bayesian network structure. In *Proc. 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*.

Singh, A. and A. Moore (2005). Finding optimal bayesian networks by dynamic programming. Technical report, Carnegie Mellon University.

Sipser, M. (2005). *Introduction to the Theory of Computation* (Second ed.). Course Technology.

Smith, A. and G. Roberts (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B 55*, 3–23.

Smith, J. (1989). Influence diagrams for statistical modeling. *Annals of Statistics 17*(2), 654–72.

Smith, J., S. Holtzman, and J. Matheson (1993). Structuring conditional relationships in influence diagrams. *Operations Research 41*(2), 280–297.

Smyth, P., D. Heckerman, and M. Jordan (1997). Probabilistic independence networks for hidden Markov probability models. *Neural Computation 9*(2), 227–269.

Sontag, D. and T. Jaakkola (2007). New outer bounds on the marginal polytope. In *Proc. 21st Conference on Neural Information Processing Systems (NIPS)*.

Sontag, D., T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss (2008). Tightening LP relaxations for MAP using message passing. In *Proc. 24th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Speed, T. and H. Kiiveri (1986). Gaussian Markov distributions over finite graphs. *The Annals of Statistics 14*(1), 138–150.

Spetzler, C. and C.-A. von Holstein (1975). Probabilistic encoding in decision analysis. *Management Science*, 340–358.

Spiegelhalter, D. and S. Lauritzen (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks 20*, 579–605.

Spiegelhalter, D. J., A. P. Dawid, S. L. Lauritzen, and R. G. Cowell (1993). Bayesian analysis in expert systems. *Statistical Science 8*, 219–283.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In *Proc. 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 491–98.

Spirtes, P., C. Glymour, and R. Scheines (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review 9*, 62–72.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, Prediction and Search*. Number 81 in Lecture Notes in Statistics. New York: Springer-Verlag.

Spirtes, P., C. Meek, and T. Richardson (1999). An algorithm for causal inference in the presence of latent variables and selection bias. See Glymour and Cooper (1999), pp. 211–52.

Srebro, N. (2001). Maximum likelihood bounded tree-width Markov networks. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Srinivas, S. (1993). A generalization of the noisy-or model. In *Proc. 9th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 208–215.

Srinivas, S. (1994). A probabilistic approach to hierarchical model-based diagnosis. In *Proc. 10th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Studený, M. and R. Bouckaert (1998). On chain graph models for description of conditional independence structures. *Annals of Statistics 26*.

Sudderth, E., A. Ihler, W. Freeman, and A. Willsky (2003). Nonparametric belief propagation. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 605–612.

Sutton, C. and T. Minka (2006). Local training and belief propagation. Technical Report MSR-TR-2006-121, Microsoft Research.

Sutton, C. and A. McCallum (2004). Collective segmentation and labeling of distant entities in information extraction. In *ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields*.

Sutton, C. and A. McCallum (2005). Piecewise training of undirected models. In *Proc. 21st Conference on Uncertainty in Artificial Intelligence (UAI)*.

Sutton, C. and A. McCallum (2007). An introduction to conditional random fields for relational learning. In L. Getoor and B. Taskar (Eds.), *Introduction to Statistical Relational Learning*. MIT

Press.

Sutton, C., A. McCallum, and K. Rohanimanesh (2007, March). Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research 8*, 693–723.

Suzuki, J. (1993). A construction of Bayesian networks from databases based on an MDL scheme. In *Proc. 9th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 266–273.

Swendsen, R. and J. Wang (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters 58*(2), 86–88.

Swendsen, R. H. and J.-S. Wang (1986, Nov). Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters 57*(21), 2607–2609.

Szeliski, R., R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother (2008, June). A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *IEEE Trans. on Pattern Analysis and Machine Intelligence 30*(6), 1068–1080. See `http://vision.middlebury.edu/MRF` for more detailed results.

Szolovits, P. and S. Pauker (1992). Pedigree analysis for genetic counseling. In *Proceedings of the Seventh World Congress on Medical Informatics (MEDINFO '92)*, pp. 679–683. North-Holland.

Tanner, M. A. (1993). *Tools for Statistical Inference*. New York: Springer-Verlag.

Tarjan, R. and M. Yannakakis (1984). Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM Journal of Computing 13*(3), 566–579.

Taskar, B., P. Abbeel, and D. Koller (2002). Discriminative probabilistic models for relational data. In *Proc. 18th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 485–492.

Taskar, B., P. Abbeel, M.-F. Wong, and D. Koller (2007). Relational Markov networks. See Getoor and Taskar (2007).

Taskar, B., V. Chatalbashev, and D. Koller (2004). Learning associative Markov networks. In *Proc. 21st International Conference on Machine Learning (ICML)*.

Taskar, B., C. Guestrin, and D. Koller (2003). Max margin Markov networks. In *Proc. 17th Conference on Neural Information Processing Systems (NIPS)*.

Tatikonda, S. and M. Jordan (2002). Loopy belief propagation and Gibbs measures. In *Proc. 18th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Tatman, J. A. and R. D. Shachter (1990). Dynamic programming and influence diagrams. *IEEE Transactions on Systems, Man and Cybernetics 20(2)*, 365–379.

Teh, Y. and M. Welling (2001). The unified propagation and scaling algorithm. In *Proc. 15th Conference on Neural Information Processing Systems (NIPS)*.

Teh, Y., M. Welling, S. Osindero, and G. Hinton (2003). Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research 4*, 1235–1260. Special Issue on ICA.

Teyssier, M. and D. Koller (2005). Ordering-based search: A simple and effective algorithm for learning bayesian networks. In *Proc. 21st Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 584–590.

Thiele, T. (1880). *Sur la compensation de quelques erreurs quasisystematiques par la methode des moindres carrees*. Copenhagen: Reitzel.

Thiesson, B. (1995). Accelerated quantification of Bayesian networks with incomplete data. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-*

95), pp. 306–311. AAAI Press.

Thiesson, B., C. Meek, D. M. Chickering, and D. Heckerman (1998). Learning mixtures of Bayesian networks. In *Proc. 14th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Thomas, A., D. Spiegelhalter, and W. Gilks (1992). BUGS: A program to perform Bayesian inference using Gibbs sampling. In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics 4*, pp. 837–842. Oxford, UK: Clarendon Press.

Thrun, S., W. Burgard, and D. Fox (2005). *Probabilistic Robotics*. Cambridge, MA: MIT Press.

Thrun, S., D. Fox, W. Burgard, and F. Dellaert (2000). Robust Monte Carlo localization for mobile robots. *Artificial Intelligence 128*(1–2), 99–141.

Thrun, S., Y. Liu, D. Koller, A. Ng, Z. Ghahramani, and H. Durrant-Whyte (2004). Simultaneous localization and mapping with sparse extended information filters. *International Journal of Robotics Research 23*(7/8).

Thrun, S., C. Martin, Y. Liu, D. Hähnel, R. Emery-Montemerlo, D. Chakrabarti, and W. Burgard (2004). A real-time expectation maximization algorithm for acquiring multi-planar maps of indoor environments with mobile robots. *IEEE Transactions on Robotics 20*(3), 433–443.

Thrun, S., M. Montemerlo, D. Koller, B. Wegbreit, J. Nieto, and E. Nebot (2004). FastSLAM: An efficient solution to the simultaneous localization and mapping problem with unknown data association. *Journal of Machine Learning Research*.

Tian, J. and J. Pearl (2002). On the testable implications of causal models with hidden variables. In *Proc. 18th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 519–527.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*(1), 267–288.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics 22*(4), 1701–1728.

Tong, S. and D. Koller (2001a). Active learning for parameter estimation in Bayesian networks. In *Proc. 15th Conference on Neural Information Processing Systems (NIPS)*, pp. 647–653.

Tong, S. and D. Koller (2001b). Active learning for structure in Bayesian networks. In *Proc. 17th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 863–869.

Torrance, G., W. Thomas, and D. Sackett (1972). A utility maximization model for evaluation of health care programs. *Health Services Research 7*, 118–133.

Tsochantaridis, I., T. Hofmann, T. Joachims, and Y. Altun (2004). Support vector machine learning for interdependent and structured output spaces. In *Proc. 21st International Conference on Machine Learning (ICML)*.

Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: Heuristics and biases. *Science 185*, 1124–1131.

van der Merwe, R., A. Doucet, N. de Freitas, and E. Wan (2000a, Aug.). The unscented particle filter. Technical Report CUED/F-INFENG/TR 380, Cambridge University Engineering Department.

van der Merwe, R., A. Doucet, N. de Freitas, and E. Wan (2000b). The unscented particle filter. In *Proc. 14th Conference on Neural Information Processing Systems (NIPS)*.

Varga, R. (2000). *Matrix Iterative Analysis*. Springer-Verlag.

Verma, T. (1988). Causal networks: Semantics and expressiveness. In *Proc. 4th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 352–359.

Verma, T. and J. Pearl (1988). Causal networks: Semantics and expressiveness. In *Proc. 4th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 69–76.

Verma, T. and J. Pearl (1990). Equivalence and synthesis of causal models. In *Proc. 6th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 255 –269.

Verma, T. and J. Pearl (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proc. 8th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 323–330.

Vickrey, D. and D. Koller (2002). Multi-agent algorithms for solving graphical games. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-02)*, pp. 345–351.

Vishwanathan, S., N. Schraudolph, M. Schmidt, and K. Murphy (2006). Accelerated training of conditional random fields with stochastic gradient methods. In *Proc. 23rd International Conference on Machine Learning (ICML)*, pp. 969–976.

Viterbi, A. (1967, April). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory 13*(2), 260–269.

von Neumann, J. and O. Morgenstern (1944). *Theory of games and economic behavior* (first ed.). Princeton, NJ: Princeton Univ. Press.

von Neumann, J. and O. Morgenstern (1947). *Theory of games and economic behavior* (second ed.). Princeton, NJ: Princeton Univ. Press.

von Winterfeldt, D. and W. Edwards (1986). *Decision Analysis and Behavioral Research*. Cambridge, UK: Cambridge University Press.

Vorobev, N. (1962). Consistent families of measures and their extensions. *Theory of Probability and Applications 7*, 147–63.

Wainwright, M. (2006). Estimating the "wrong" graphical model: Benefits in the computation-limited setting. *Journal of Machine Learning Research 7*, 1829–1859.

Wainwright, M., T. Jaakkola, and A. Willsky (2003a). Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Transactions on Information Theory 49*(5).

Wainwright, M., T. Jaakkola, and A. Willsky (2003b). Tree-reweighted belief propagation and approximate ML estimation by pseudo-moment matching. In *Proc. 9thWorkshop on Artificial Intelligence and Statistics*.

Wainwright, M., T. Jaakkola, and A. Willsky (2004, April). Tree consistency and bounds on the performance of the max-product algorithm and its generalizations. *Statistics and Computing 14*, 143–166.

Wainwright, M., T. Jaakkola, and A. Willsky (2005). MAP estimation via agreement on trees: Message-passing and linear programming. *IEEE Transactions on Information Theory*.

Wainwright, M., T. Jaakkola, and A. S. Willsky (2001). Tree-based reparameterization for approximate estimation on loopy graphs. In *Proc. 15th Conference on Neural Information Processing Systems (NIPS)*.

Wainwright, M., T. Jaakkola, and A. S. Willsky (2002a). Exact map estimates by (hyper)tree agreement. In *Proc. 16th Conference on Neural Information Processing Systems (NIPS)*.

Wainwright, M., T. Jaakkola, and A. S. Willsky (2002b). A new class of upper bounds on the log partition function. In *Proc. 18th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Wainwright, M. and M. Jordan (2003). Graphical models, exponential families, and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley.

Wainwright, M. and M. Jordan (2004). Semidefinite relaxations for approximate inference on graphs with cycles. In *Proc. 18th Conference on Neural Information Processing Systems (NIPS)*.

Wainwright, M., P. Ravikumar, and J. Lafferty (2006). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Proc. 20th Conference on Neural Information*

Processing Systems (NIPS).

Warner, H., A. Toronto, L. Veasey, and R. Stephenson (1961). A mathematical approach to medical diagnosis — application to congenital heart disease. *Journal of the American Madical Association 177*, 177–184.

Weiss, Y. (1996). Interpreting images by propagating bayesian beliefs. In *Proc. 10th Conference on Neural Information Processing Systems (NIPS)*, pp. 908–914.

Weiss, Y. (2000). Correctness of local probability propagation in graphical models with loops. *Neural Computation 12*, 1–41.

Weiss, Y. (2001). Comparing the mean field method and belief propagation for approximate inference in MRFs. In M. Opper and D. Saad (Eds.), *Advanced mean field methods*, pp. 229–240. Cambridge, Massachusetts: MIT Press.

Weiss, Y. and W. Freeman (2001a). Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation 13*.

Weiss, Y. and W. Freeman (2001b). On the optimality of solutions of the max-product belief propagation algorithm in arbitrary graphs. *IEEE Transactions on Information Theory 47*(2), 723–735.

Weiss, Y., C. Yanover, and T. Meltzer (2007). MAP estimation, linear programming and belief propagation with convex free energies. In *Proc. 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*.

Welling, M. (2004). On the choice of regions for generalized belief propagation. In *Proc. 20th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Welling, M., T. Minka, and Y. Teh (2005). Structured region graphs: Morphing EP into GBP. In *Proc. 21st Conference on Uncertainty in Artificial Intelligence (UAI)*.

Welling, M. and S. Parise (2006a). Bayesian random fields: The Bethe-Laplace approximation. In *Proc. 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*.

Welling, M. and S. Parise (2006b). Structure learning in Markov random fields. In *Proc. 20th Conference on Neural Information Processing Systems (NIPS)*.

Welling, M. and Y.-W. Teh (2001). Belief optimization for binary networks: a stable alternative to loopy belief propagation. In *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Wellman, M. (1985). Reasoning about preference models. Technical Report MIT/LCS/TR-340, Laboratory for Computer Science, MIT.

Wellman, M., J. Breese, and R. Goldman (1992). From knowledge bases to decision models. *Knowledge Engineering Review 7*(1), 35–53.

Wellman, M. and J. Doyle (1992). Modular utility representation for decision-theoretic planning. In *Procec. First International Conference on AI Planning Systems*, pp. 236–42. Morgan Kaufmann.

Wellman, M. P. (1990). Foundamental concepts of qualitative probabilistic networks. *Artificial Intelligence 44*, 257–303.

Wellner, B., A. McCallum, F. Peng, and M. Hay (2004). An integrated, conditional model of information extraction and coreference with application to citation matching. In *Proc. 20th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 593–601.

Wermuth, N. (1980). Linear recursive equations, covariance selection and path analysis. *Journal of the American Statistical Association 75*, 963–975.

Werner, T. (2007). A linear programming approach to max-sum problem: A review. *IEEE Trans. on Pattern Analysis and Machine Intelligence 29*(7), 1165–1179.

West, M. (1993). Mixture models, Monte Carlo, Bayesian updating and dynamic models. *Comput-*

ing Science and Statistics 24, 325–333.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester, United Kingdom: John Wiley and Sons.

Wiegerinck, W. (2000). Variational approximations between mean field theory and the junction tree algorithm. In *Proc. 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 626–636.

Wold, H. (1954). Causality and econometrics. *Econometrica 22*, 162–177.

Wood, F., T. Griffiths, and Z. Ghahramani (2006). A non-parametric bayesian method for inferring hidden causes. In *Proc. 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 536–543.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research 20*, 557–85.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics 5*, 161–215.

Xing, E., M. Jordan, and S. Russell (2003). A generalized mean field algorithm for variational inference in exponential families. In *Proc. 19th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 583–591.

Yanover, C., T. Meltzer, and Y. Weiss (2006, September). Linear programming relaxations and belief propagation — an empirical study. *Journal of Machine Learning Research 7*, 1887–1907.

Yanover, C., O. Schueler-Furman, and Y. Weiss (2007). Minimizing and learning energy functions for side-chain prediction. In *Proc. International Conference on Research in Computational Molecular Biology (RECOMB)*, pp. 381–395.

Yanover, C. and Y. Weiss (2003). Finding the M most probable configurations using loopy belief propagation. In *Proc. 17th Conference on Neural Information Processing Systems (NIPS)*.

Yedidia, J., W. Freeman, and Y. Weiss (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Information Theory 51*, 2282–2312.

Yedidia, J. S., W. T. Freeman, and Y. Weiss (2000). Generalized belief propagation. In *Proc. 14th Conference on Neural Information Processing Systems (NIPS)*, pp. 689–695.

York, J. (1992). Use of the Gibbs sampler in expert systems. *Artificial Intelligence 56*, 115–130.

Yuille, A. L. (2002). CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. *Neural Computation 14*, 1691–1722.

Zhang, N. (1998). Probabilistic inference in influence diagrams. In *Proc. 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 514–522.

Zhang, N. and D. Poole (1994). A simple approach to Bayesian network computations. In *Proceedings of the 10th Biennial Canadian Artificial Intelligence Conference*, pp. 171–178.

Zhang, N. and D. Poole (1996). Exploiting contextual independence in probabilistic inference. *Journal of Artificial Intelligence Research 5*, 301–328.

Zhang, N., R. Qi, and D. Poole (1993). Incremental computation of the value of perfect information in stepwise-decomposable influence diagrams. In *Proc. 9th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 400–407.

Zhang, N. L. (2004). Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research 5*, 697–723.

Zoeter, O. and T. Heskes (2006). Deterministic approximate inference techniques for conditionally Gaussian state space models. *Statistical Computing 16*, 279–292.

Zweig, G. and S. J. Russell (1998). Speech recognition with dynamic Bayesian networks. In *Proc. 14th Conference on Artificial Intelligence (AAAI)*, pp. 173–180.

Notation Index

$|A|$ — Cardinality of the set A, 20
$\phi_1 \times \phi_2$ — Factor product, 107
$\gamma_1 \bigoplus \gamma_2$ — Joint factor combination, 1104
$p(\boldsymbol{Z}) \bigoplus g(\boldsymbol{Z})$ — Marginal of $g(\boldsymbol{Z})$ based on $p(\boldsymbol{Z})$, 631
$\sum_Y \phi$ — Factor marginalization, 297
$X \rightleftharpoons Y$ — Bi-directional edge, 34
$X \rightarrow Y$ — Directed edge, 34
$X — Y$ — Undirected edge, 34
$X \leftrightarrow Y$ — Non-ancestor edge (PAGs), 1049
$X \circ\!\!\rightarrow Y$ — Ancestor edge (PAGs), 1049
$\langle x, y \rangle$ — Inner product of vectors x and y, 262
$\|P - Q\|_1$ — L_1 distance, 1143
$\|P - Q\|_2$ — L_2 distance, 1143
$\|P - Q\|_\infty$ — L_∞ distance, 1143
$(\boldsymbol{X} \perp \boldsymbol{Y})$ — Independence of random variables, 24
$(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$ — Conditional independence of random variables, 24
$(\boldsymbol{X} \perp_c \boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{c})$ — Context-specific independence, 162

$\boldsymbol{I}\{\cdot\}$ — Indicator function, 32

$\mathcal{A}(\boldsymbol{x} \rightarrow \boldsymbol{x}')$ — Acceptance probability, 517
$\aleph$ — Template attributes, 214
$\alpha(A)$ — The argument signature of attribute A, 213
$Ancestors_X$ — Ancestors of X (in graph), 36
argmax, 26
A — A template attribute, 213

$Beta(\alpha_1, \alpha_0)$ — Beta distribution, 735
β_i — Belief potential, 352
$\mathcal{B}_{\mathcal{I}[\sigma]}$ — Induced Bayesian network, 1093
$\mathcal{B}$ — Bayesian network, 62
$\mathcal{B}_0$ — Initial Bayesian network (DBN), 204
$\mathcal{B}_\rightarrow$ — Transition Bayesian network (DBN), 204
$\mathcal{B}_{\boldsymbol{Z}=\boldsymbol{z}}$ — Mutilated Bayesian network, 499
$\mathrm{Boundary}_X$ — Boundary around X (in graph), 34

$\mathcal{C}(K, \boldsymbol{h}, g)$ — Canonical form, 609
$\mathcal{C}(\boldsymbol{X}; K, \boldsymbol{h}, g)$ — Canonical form, 609
$\mathcal{C}[v]$ — Choices, 1085
Ch_X — Children of X (in graph), 34
$\boldsymbol{C}_i$ — Clique, 346
$\boldsymbol{x} \sim \boldsymbol{c}$ — Compatability of values , 20
$\mathrm{cont}(\gamma)$ — Joint factor contraction, 1104
$\boldsymbol{Cov}[X; Y]$ — Covariance of X and Y, 248

$\boldsymbol{D}$ — A subclique, 104
Δ — Discrete variables (hybrid models), 605
$\boldsymbol{d}$ — Value of a subclique, 104
$\mathcal{D}^+$ — Complete data, 871
$\mathcal{D}$ — Empirical samples (data), 698
$\mathcal{D}$ — Sampled data, 489
$\mathcal{D}^*$ — Complete data, 912
$\mathcal{D}$ — Decisions, 1089
$\mathrm{Descendants}_X$ — Descendants of X (in graph), 36
$\tilde{\delta}_{i \rightarrow j}$ — Approximate sum-product message, 435
$\delta_{i \rightarrow j}$ — Sum-product message, 352
$\mathrm{Dim}[\mathcal{G}]$ — Dimension of a graph, 801
$Dirichlet(\alpha_1, \ldots, \alpha_K)$ — Dirichlet distribution, 738
$\boldsymbol{D}(P \| Q)$ — Relative entropy, 1141
$\boldsymbol{D}_{var}(P; Q)$ — Variational distance, 1143
$\mathbf{Down}^*(r)$ — Downward closure, 422
$\mathbf{Down}^+(r)$ — Extended downward closure, 422
$\mathbf{Down}(r)$ — Downward regions, 422
$do(Z := z), do(z)$ — Intervention, 1010

$d\text{-}sep_{\mathcal{G}}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z})$ — d-separation, 71

$\mathcal{E}$ — Edges in MRF, 127
$\mathrm{EU}[\mathcal{D}[a]]$ — Expected utility, 1061
$\mathrm{EU}[\mathcal{I}[\sigma]]$ — Expected utility of σ, 1093
$\hat{\boldsymbol{E}}_{\mathcal{D}}(f)$ — Empirical expectation, 490
$\boldsymbol{E}_{\mathcal{D}}[f]$ — Empirical expectation, 700
$\boldsymbol{E}_P[X]$ — Expectation (mean) of X, 31
$\boldsymbol{E}_P[X \mid \boldsymbol{y}]$ — Conditional expectation, 32
$\boldsymbol{E}_{X\sim P}[\cdot]$ — Expectation when $X \sim P$, 387

$f(\boldsymbol{D})$ — A feature, 124
$F[\tilde{P}, Q]$ — Energy functional, 385, 881
$\tilde{F}[\tilde{P}_\Phi, \boldsymbol{Q}]$ — Region Free Energy functional, 420
$\tilde{F}[\tilde{P}_\Phi, \boldsymbol{Q}]$ — Factored energy functional, 386
$\mathrm{FamScore}(X_i \mid \mathrm{Pa}_{X_i} : \mathcal{D})$ — Family score, 805
$\mathcal{F}$ — Feature set, 125
$\mathcal{F}$ — Factor graph, 123

$\mathcal{G}$ — Directed graph, 34
$\boldsymbol{\mathcal{G}}$ — Partial ancestral graph, 1049
Γ — Continuous variables (hybrid models), 605
γ — Template assignment, 215
$\mathrm{Gamma}(\alpha, \beta)$ — Gamma distribution, 900
$\Gamma(x)$ — Gamma function, 736

$\mathcal{H}$ — Missing data, 859
$\mathcal{H}$ — Undirected graph, 34
$\boldsymbol{H}_P(X)$ — Entropy, 1138
$\boldsymbol{H}_P(X \mid Y)$ — Conditional entropy, 1139
$\tilde{\boldsymbol{H}}^{\boldsymbol{\kappa}}_{\boldsymbol{Q}}(\mathcal{X})$ — Weighted approximate entropy, 415

$\mathcal{I}$ — Influence diagram, 1090
$\mathcal{I}(\mathcal{G})$ — Markov independencies of $\mathcal{G}$, 72
$\mathcal{I}_\ell(\mathcal{G})$ — Local Markov independencies of $\mathcal{G}$, 57
$\mathcal{I}(P)$ — The independencies satisfied by P, 60
$\boldsymbol{I}_P(X;Y)$ — Mutual infromation, 1140
$Interface_{\mathcal{H}}(\boldsymbol{X};\boldsymbol{Y})$ — $\boldsymbol{Y}$-interface of $\boldsymbol{X}$, 464

$\mathcal{J}$ — Lagrangian, 1168
J — Precision matrix, 248

$\mathcal{K}$ — Partially directed graph, 34
$\mathcal{K}^+[\boldsymbol{X}]$ — Upward closed subgraph, 35
κ — Object skeleton (template models), 214
κ_r — Counting number of region r, 415
$\boldsymbol{K}_i$ — Member of a chain, 37
$\mathcal{K}[\boldsymbol{X}]$ — Induced subgraph, 35

$\ell_{\mathrm{PL}}(\boldsymbol{\theta} : \mathcal{D})$ — Pseudolikelihood, 970
$L(\theta : \mathcal{D})$ — Likelihood function, 721
$Local[\mathcal{U}]$ — Local polytope, 412
$\ell(\hat{\boldsymbol{\theta}}_{\mathcal{G}} : \mathcal{D})$ — Maximum likelihood value, 791
$\ell(\theta : \mathcal{D})$ — Log-likelihood function, 719
$\ell_{\boldsymbol{Y}\mid\boldsymbol{X}}(\boldsymbol{\theta} : \mathcal{D})$ — Conditional log-likelihood function, 951
$loss(\xi : \mathcal{M})$ — Loss function, 699

$\mathcal{M}^*$ — Model that generated the data, 698
$\text{M-project-distr}_{i,j}$ — M-projection, 436
$M[\boldsymbol{x}]$ — Counts of event $\boldsymbol{x}$ in data, 724
$Marg[\mathcal{U}]$ — Marginal polytope, 411
$marg_{\boldsymbol{W}}(\gamma)$ — Joint factor marginalization, 1104
$MaxMarg_f(\boldsymbol{x})$ — Max marginal of f, 553
$\mathcal{M}[\mathcal{G}]$ — Moralization of $\mathcal{G}$, 134
$\mathcal{M}$ — A model, 699
$\bar{M}_{\boldsymbol{\theta}}[\boldsymbol{x}]$ — Expected counts, 871
$\tilde{\mathcal{M}}$ — Learned/estimated model, 698

$\mathcal{N}\left(\mu; \sigma^2\right)$ — A Gaussian distribution, 28
$\mathcal{N}\left(X \mid \mu; \sigma^2\right)$ — Gaussian distribution over X, 616
Nb_X — Neighbors of X (in graph), 34
$\mathrm{NonDescendants}_X$ — Non-descendants of X (in graph), 36
$\mathcal{NP}$, 1151

$\mathcal{O}$ — Outcome space, 1060
$O(f(\cdot))$ — "Big O" of f, 1148
$\mathcal{O}^\kappa[\mathsf{Q}]$ — Objects in κ (template models), 214

$\mathcal{P}$, 1151
$P(X \mid Y)$ — Conditional distribution, 22
$P(x), P(x, y)$ — Shorthand for $P(X = x)$, $P(X = x, Y = y)$, 21
P^* — Distribution that generated the data, 698
$P \models \ldots$ — P satisfies ..., 23
Pa_X — Parents of X (in graph), 34
pa_X — Value of Pa_X, 157
$\mathrm{Pa}^{\mathcal{G}}_{X_i}$ — Parents of X_i in $\mathcal{G}$, 57
$\hat{P}_{\mathcal{D}}(A)$ — Empirical distribution, 703
$\hat{P}_{\mathcal{D}}(\boldsymbol{x})$ — Empirical distribution, 490
$\boldsymbol{\theta}$ — Parameters, 262, 720
$\hat{\boldsymbol{\theta}}$ — MLE parameters, 726
ϕ — A factor (Markov network), 104

$\phi[\boldsymbol{U} = \boldsymbol{u}]$ — Factor reduction, 110
π — Lottery, 1060
$\pi(\boldsymbol{X})$ — Stationary probability, 509
$\tilde{P}_\Phi(\mathcal{X})$ — Unnormalized measure defined by Φ, 345
$\psi_i(\boldsymbol{C}_i)$ — Initial potential, 349
$\tilde{P}$ — Learned/estimated distribution, 698

$\boldsymbol{Q}$ — Approximating distribution, 383
$\mathcal{Q}$ — Template classes, 214

$\mathcal{R}$ — Region graph, 419
$\mathbb{R}$ — Real numbers, 27
ρ — A rule, 166
$\mathcal{R}$ — Rule set, 168

$\mathcal{S}$ — Event space, 15
σ — Std of a Gaussian distribution, 28
σ — Strategy, 1092
$\sigma^{(t)}(\cdot)$ — Belief state, 652
Scope$[\phi]$ — Scope of a factor, 104
$\text{score}_B(\mathcal{G} : \mathcal{D})$ — Bayesian score, 795
$\text{score}_{BIC}(\mathcal{G} : \mathcal{D})$ — BIC score, 802
$\text{score}_{CS}(\mathcal{G} : \mathcal{D})$ — Cheeseman-Stutz score, 913
$\text{score}_L(\mathcal{G} : \mathcal{D})$ — Likelihood score, 791
$\text{score}_{L_1}(\boldsymbol{\theta} : \mathcal{D})$ — L_1 score, 988
$\text{score}_{Laplace}(\mathcal{G} : \mathcal{D})$ — Laplace score, 910
$\text{score}_{\text{MAP}}(\boldsymbol{\theta} : \mathcal{D})$ — MAP score, 898
$sep_{\mathcal{H}}(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})$ — Separation in $\mathcal{H}$, 114
$\text{sigmoid}(x)$ — Sigmoid function, 145
$\boldsymbol{S}_{i,j}$ — Sepset, 140, 346
$succ(v, c)$ — Successor (decision trees), 1085

$\mathcal{T}$ — Clique tree, 140, 347
Υ — Template clique tree, 656
$\mathcal{T}$ — Decision tree, 1085
$\mathsf{t}(\boldsymbol{\theta})$ — Natural parameters function, 261
$\tau(\xi)$ — Sufficient statistics function, 261, 721
Θ — Parameter space, 261, 720
$\mathcal{T}(\boldsymbol{x} \to \boldsymbol{x}')$ — Transition probability, 507

$\mathcal{U}$ — Cluster graph, 346
$\mathcal{U}$ — Response variables, 1029
μ — Mean of a Gaussian distribution, 28
$U(o)$ — Utility function, 1060
$\mu_{i,j}$ — Sepset beliefs, 358
$\text{Unif}[\text{a}, \text{b}]$ — Uniform distribution on $[a, b]$, 28
$\mathbf{Up}^*(r)$ — Upward closure, 422
$\mathbf{Up}(r)$ — Upward regions, 422
$\mathcal{U}$ — Utility variables, 1090
U^X — Response variable, 1029

$Val(X)$ — Possible values of X, 20
$\mathbb{V}ar_P[X]$ — Variance of X, 33
$\text{VPI}_{\mathcal{I}}(D \mid X)$ — Value of perfect information, 1122
$\nu_r, \nu_i, \nu_{r,i}$ — Convex counting numbers, 416

$\boldsymbol{W}_{<(i,j)}$, 348

$\mathcal{X}$ — The set of all variables in the domain, 21
ξ — An assignment to $\mathcal{X}$, 79
X, Y, Z — Random variables, 20
$\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ — Random variable sets, 20
$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ — Values of random variable sets, 20
x^0, x^1 — False/True values of X, 20
$\boldsymbol{x}\langle\boldsymbol{Y}\rangle$ — Assignment in $\boldsymbol{x}$ to variables in $\boldsymbol{Y}$, 21
x[m]$\boldsymbol{x}[m]$ — m'th data instance (i.i.d. samples), 698
x^i — The i'th value of X, 20
$\mathcal{X}_\kappa[A]$ — Ground random variables, 214
$\xi[m]$ — m'th data instance (i.i.d. samples), 488
ξ^{map} — MAP assignment, 552
$X^{(t)}$ — X at time t, 200
$X^{(t_1:t_2)}$ — X in the interval $[t_1, t_2]$, 200
$X \sim \ldots$ — X is distributed according to ..., 28

Z — Partition function, 105

Subject Index

2-TBN, 202
3-SAT, 288, 1151

abduction, 1134
action, 1061
  joint, 1117
active learning, 1055
activity recognition, 952
Algorithm
  Alpha-Expand, 593, *593*
  Alpha-Expansion, *593*
  BU-Message, 367, *367*, 368, 440, 441
  BU-message, 400
  Beam-Search, *1158*
  Branch-and-Bound, 1161, *1161*
  Build-Minimal-I-Map, 80, *80*, 142, 786
  Build-PDAG, *89*, 90–92, 786, 787, 790, 839, 843, 1042, 1043
  Build-PMAP-Skeleton, 787
  Build-PMap-Skeleton, *85*, 86, 89, 90, 101, 787, 980, 1005, 1051
  Build-Saturated-Region-Graph, *423*
  CGraph-BU-Calibrate, 398, 400, 413
  CGraph-SP-Calibrate, *397*, 413, 428
  CLG-M-Project-Distr, *622*, 628
  CSI-sep, *173*
  CTree-BU-Calibrate, *367*, 398, 628
  CTree-BU-calibrate, 391
  CTree-Filter-DBN, *657*
  CTree-Query, *371*
  CTree-SP-Calibrate, *357*, 364, 365, 368, 398, 413, 436
  CTree-SP-Upward, *353*, 378, 612
  CTree-SP-calibrate, 388, 413
  Compute-ESS, 873, *873*, 938
  Compute-Gradient, 867, *867*
  Cond-Prob-VE, *304*, 317
  Conditioning, 604
  Conjugate-Gradient-Ascent, *1167*
  Convex-BP-Msg, *418*
  Cross-Validation, *707*
  DP-Merge-Split-Proposal, *942*
  Data-Dependent-LW, 502, *502*, 504
  EP-Message, 440, *441*, 443, 628
  Estimate-Parameters, 922, 941
  Evaluate, 707, *707*
  Expectation-Maximization, *873*, 922
  Factor-Product, *359*
  Factored-Project, *434*, 435
  Fibonacci, 1150, *1150*
  Find-Immoralities, 89
  Forward-Sample, *489*
  Generalized-MP-BP, *573*
  Generalized-VE-for-IDs, *1105*, 1106, 1107
  Gibbs-Sample, *506*
  Gradient-Ascent, *1164*
  Greedy-Local-Search, 815, *1155*, 1156
  Greedy-MN-Structure-Search, *986*, 990, 992
  Greedy-Ordering, *314*, 340
  Holdout-Test, *707*
  Incremental-E-Step, 939, *939*
  Incremental-EM, *939*
  Initialize-CGraph, 397, *397*, 573
  Initialize-CTree, 367, *367*
  Initialize-Cliques, 353, *353*, 357
  Iterated-Optimization-for-IDs, 1116, *1116*, 1131
  K-Best, 1158
  LW-2TBN, 666, *666*, 670
  LW-DBN, *666*
  LW-Sample, 493, *493*, 502
  LearnProc, 706, 707
  LegalOp, 1157, *1157*, 1158
  M-Project-Distr, *443*, 621

MCMC-Sample, *509*
MEU-for-Decision-Trees, *1088*, 1098
Mark-Immoralities, *86*, 87, 89, 102, 787
Max-Cardinality, 312, *312*, 313
Max-Message, *562*
Max-Product-Eliminate-Var, 557, *557*
Max-Product-VE, *557*
Max-Weight-Spanning-Tree, *1147*
Mean-Field, 455, *455*, 459
MinCut-MAP, *591*, 593
MinCut, 591
Msg-Truncated-1-Norm, *603*
Parameter-Optimize, 986
Particle-Filter-DBN, *670*
Perturb, 817, 818
Proposal-Distribution, *941*
Reachable, *75*, 76, 102
Rule-Split, 332, *332*, 333
Rule-Sum-Product-Eliminate-Var, *333*, 601
SP-Message, 353, *353*, 357, 368, 378, 397, *397*, 407, 437, 567, 612
Search-with-Data-Perturbation, *817*
Search-with-Restarts, *1159*
Search, 817, 1159
Structural-EM, *922*
Structure-Learn, 922
Sum-Product-Conditioning, *317*
Sum-Product-Eliminate-Var, 298, *298*, 306, 347, 611
Sum-Product-VE, *298*, 299, 304, 313, 331, 371, 611
Tabu-Structure-Search, *1157*
Topological-Sort, *1146*
Traceback-MAP, 557, *557*, 558, 561, 601
Train-And-Test, 707, *707*
alignment, *see* correspondence
alpha-beta swap, 592, 602
alpha-expansion, 592
ancestor, 36
argument, 213
factor, 216
feature, 229
signature, 213, 223
parent, 221, 223
assignment
local optimality, 566–567, 569
MAP, 26, 967
strong local maximum, 570–572, 602
attribute, 213
object-valued, 234
average causal effect, 1032

back-door
criterion, 1020–1021
trail, 1020
backward induction, 1098
decision tree, 1087
bag of words, 766
barren node, 98, 136
basin flooding, 816, 1156
Bayes' rule, 18
BayesBall, 94
Bayesian classifier, 727
Bayesian estimation, 735, 739, 752, 781, 782, 824
Bayesian networks, 741–750
BDE prior, *see* BDe prior
BGe prior, *see* BGe prior
Dirichlet prior, 739, 740
Gaussian, 779–780
incomplete data, 898–908, 1052
MCMC, 899–904
variational, *see* variational Bayes
nonparametric, 730–731, 928–930
shared parameters, 762–763
Bayesian model averaging, 785, 824–832, 928, 1043
computational complexity, 827
MCMC, 829–832
Bayesian network, 5, 62
conditional, *see* conditional Bayesian network
gradient, 339, 483
structure, 57
Bayesian score, 983
BDe prior, 749, 806, 835, 844, 848
shared parameters, 780
beam search, 890
belief propagation
asynchronous, 408, 417
clique tree, 355–358
cluster graph, 396–399
convergence, 392, 401–403, 407–411, 417–419
convergence point, 412–413, 479
stability, 408, 413
convex, 416–419
damping, 408, 479
EM, 897
frustrated loop, 568
Gaussian, *see* Gaussian, belief propagation

local maxima, 409
loopy, 393, 405, 962
Markov network learning, 963–965
max-product, 562, 593, 602
convergence, 602
message scheduling, 408
nonparametric, 646, 649
operator, 402
region graph, 423–428
residual, 408
sum-product, 356
synchronous, 402, 408
tree reparameterization, 408
tree-CPDs, 478
tree-reweighted, 418, 576, 593, 968
belief state, 652
prior, 653
projection, 663
reduced, 656
beliefs, 358
Beta distribution, 735–737
BGe prior, 840
bias, 710
bias-variance trade-off, 704
bigram model, 764
bipartite matching, 534
BK algorithm, 690
BN2O network, 177, 197
Boltzmann distribution, 126
Bonferroni correction, 843
bootstrap, 1046
bow pattern, 1024
BUGS system, 525–526, 543

c-separation, 150, 156
CAI-map, 1076
minimal, 1077
perfect, 1077
calibrated, 358
CAMEL, 964, 1004
canonical form, 609, 649
division, 610
marginalization, 610
well-defined, 611
operations, 610–611
product, 610
reduction, 611
vacuous, 610
canonical table, 618
marginalization, 619–621
weak, 620
operations, 618–621
causal
effect, 1014
independence, 1056
mechanism, 175, 1014
model, 1014–1030
augmented, 1017–1020, 1022–1024
functional, 1029–1030
identifiability, 1042
causal Markov assumption, 1041
causal model learning, 1040–1053
Bayesian model averaging, 1043
constraint-based, 1042–1043
functional causal model, 1051–1053, 1056
interventional data, 1044–1047
latent variables, 1048–1051
constraint-based, 1048–1051, 1056
score-based, 1048
cellular network reconstruction, 1046–1047
central limit theorem, 1144
Markov chain, 521
certainty equivalent, 1066
chain component, 37, 148
chain graph, 37, 148
c-separation, *see* c-separation
distribution, 149
model, 148
chain rule
Bayesian networks, 54, 62
conditional probabilities, 18, 47
entropy, 1139
mutual information, *see* mutual information, chain rule
relative entropy, 1142
chance variable, 1089
Chebyshev's inequality, 33
Cheeseman-Stutz score, *see* marginal likelihood approximation, Cheeseman-Stutz
Chernoff bound, 491, 501, 1145
χ^2
distribution, 790
statistic, 788, 843, 848
child, 34
Chinese restaurant process, 930
chordal graph, 311
clarity test, 64
class, 213

classification, 50, 727
collective, 952
error, 701
task, 700
text, 766
CLG, *see* Gaussian, conditional linear
CLG network, 190, 645, 684
computational complexity, 615–617
clique, 35
clique potentials, 109
clique tree, 140, 346–348, 481, 549, 673, 937
algorithm
correctness, 353–355
beliefs, 352, 357, 365
calibrated, 355–358, 384
CLG network, 626–630
clique, 348
downstream, 347
initial potential, 351
ready, 350, 356
upstream, 347
computational complexity, 358, 374
construction, 372–376, 379, 380
downward pass, 356, 655
family preservation, *see* cluster graph, family preservation
incremental update, 369–370, 379
inference as optimization, 387–390
influence diagram, 1109, 1117, 1131, 1132
invariant, 361–363, 368
max-product, 564, 568
max-calibrated, 563
max-product, 562–565
traceback, 566
measure, 361–364, 383–384, 564
message, 345
scheduling, 357
message passing, *see* message passing
multiple queries, 371–372
nested, 377
out-of-clique inference, 370–371, 379
reparameterization, 362
rule-based CPDs, 379
running intersection property, 347–348, 353
sampling, 544
sepset, 140
strong root, 627
structure changes, 378, 379
sum-product, 352
template, 656
upward pass, 356, 378, 654
cluster graph, 346, 396
Bayesian network, 478
beliefs, 396
low-temperature-limit, 583
Bethe, 405, 414, 415, 573
calibrated, 396–398, 412
construction, 404–411
family preservation, 346, 420
induced subgraph, 570
invariant, 399–400
max-calibrated, 583
message passing, *see* message passing
out-of-cluster inference, 481
residual, 401, 477
running intersection property, 396, 407
sepset, 346
template, 664
tree consistency, 401
clustering, 875
Bayesian, *see* naive Bayes, clustering, 875, 902–908, 915–916
collaborative filtering, 823, 877
collapsed sampling, *see* Gibbs, collapsed, *see* importance sampling, collapsed, *see* MCMC, collapsed, 526–532, 645, 650
compression, 1137
computational complexity, 1147–1149
asymptotic, 1147
running time, 1148
theory, 1150
concentration phenomenon, 777
condensation, *see* filter, particle
conditional Bayesian network, 191
conditional covariance, 259
conditional expectation, 32, 451
conditional independence, *see* independence
conditional preference structure, 1072
conditional probability, 18
conditional probability distribution, *see* CPD
conditional probability table, *see* table-CPD
conditional random field, 113, 143, 191, 197, 710, 950, 952
linear-chain, 146
skip-chain, 146
conditioning, 315–325
bounded, 540
computational complexity, 320–325

cutset, 318
incremental, 540
induced graph, 322
marginal MAP, 604
rule-based CPDs, 334
confidence interval, 719
confounding factor, 1012–1014
constraint, 388
equality, 1168
expectation consistency, 446, 447
local polytope, *see* local polytope
marginal consistency, 384, 387, 416
region graph, 421
marginal polytope, *see* marginal polytope
mean field, 455
constraint generation, 976, 1005
constraint propagation, 89
constraint satisfaction problem, 569
context-specific independence, *see* independence, context-specific
contingency table, 152
contingent dependency model, 223
contraction, 402
contrastive
divergence, 974–975
objective, 970
convergence bound, 489, 771, 1145–1146
convergence rate, 888
convex optimization, 976
coordinate ascent, 881
coordination graph, 1117
correspondence, 165, 236, 532–536, 544, 550
correlated, 535
EM, 534
Metropolis Hastings, 534
mutual exclusion, 533–534
variable, 166, 533, 893
counterfactual
query, 1010, 1026–1027, 1034–1040
twinned network, 1035–1037
world, 1034
counterfactual twinned network, 1125
counting numbers, 415, 420, 573
convex, 416, 419, 574
CPD, 47, 53, 62
aggregator, 225, 245
conditional linear Gaussian, 190, 618
decomposition
causal independence, 325–329
context-specific independence, 341
deterministic, 158
encapsulated, 192
Gaussian, *see* Gaussian, linear
linear Gaussian, 187
logistic, 145, 179, 197, 225, 483
multinomial, 181, 970
multiplexer, 165
noisy-and, 196
noisy-max, 183, 196
noisy-or, 176, 196, 197, 225, 936, 1037
requisite, 100, 1018, 1112
rule-based, 168, 195, 601
inference, *see* variable elimination, rule-based CPDs
table-CPD, 157, 725
tree-CPD, 164, 195, 196
CPT, *see* CPD, table
CRF, *see* conditional random field
cross-validation, 706, 844, 960
CSI-separation, 173, 196
computational complexity, 196
cycle, 37
cyclic graphical model, 95

d-separation, 71
completeness, 72
soundness, 72
DAG, 37, 57
data
complete, 712
completion, 869, 881, 912, 921
incomplete, 712, 849
interventional, 1040, 1044, 1056
observability, 712
observational, 1040
weighted, 817, 870
data association, *see* correspondence, 165, 244, 532, 550, 680, 893, 940
data fragmentation, 726, 784
data imputation, 869
data perturbation, 816
data-driven approach, 6
decision diagram, 170
decision rule, 1091
deterministic, 1091
fully mixed, 1111
locally optimal, 1109, 1110
optimization, 1107, 1108, 1130

iterated, 1115–1117, 1131
local, 1111
decision theory, 1059, 1068
decision tree, 1085, 1096–1097
strategy, 1087
decision variable, 1017, 1089
decision-making situation, 1061
declarative representation, 1, 1133
deep belief networks, 1000
degree, 34
bounded, 992
density estimation, 699, 784
density function, 27–31
conditional, 31
joint, 29
dependency network, 96, 822, 823
descendant, 36
detailed balance, 515, 546
deterministic separation, 160
digamma function, 907
directed acyclic graph, *see* DAG
Dirichlet distribution, *see* BDe prior, 738, 746–750
mixture, 779
posterior, 738
sampling, 900
variational update, 906–907
Dirichlet process, 929–930, 941–942
discretization, 606
discriminative training, 709, 950, 997
distance measure, 1140–1143
distance metric, 1140, 1143
distribution, 16
Bernoulli, 20
conditional, 22
cumulative, 28
empirical, 703
Gamma, 765, 780, 900
Gaussian, *see* Gaussian, 720
joint, 3, 21
Laplacian, 959
marginal, 21
mixture, *see* Gaussian, mixture, 484, 713, 875, 915
multinomial, 20, 720
normal-Gamma, 751
Poisson, 283
positive, 25, 116
posterior, 3
prior, 47
support, 494
uniform, 28
duality, 957, 1171–1172
convex, 470
dynamic Bayesian network, 202–205, 837
fully persistent, 658
parameter estimation, 781
structure learning, 846
dynamic programming, 292–296, 337, 356, 371, 482, 596, 1149
dynamical system, *see* filter
continuous time Bayesian network, 242
Dynamic Bayesian network, *see* dynamic Bayesian network
hidden Markov model, *see* hidden Markov model
linear, 211
Markovian, 201
semi-Markov, 202, 243
semi-Markovian, 243
stationary, 202
switching linear, 212, 684

E-step, 872, 874, 907
variational, 896
edge
covered, 78, 100
covering, 78
directed, 34
fill, 307, 340
inter-time-slice, 204
intra-time-slice, 204
reversal, 78, 99, 545, 673
spurious, 173
undirected, 34
EKF, *see* Kalman filter, extended
EM, 535, 907
accelerated, 892
approximate inference, 893–897
Bayesian network, 868–897
computational complexity, 891
table-CPD, 872–874
belief propagation, 897
clustering, 875–877
convergence, 887
practice, 885–887, 890–892
theory, 874–875, 877–884
dynamic Bayesian network, 937

exponential family, 874
hard assignment, 876, 884–885, 889, 937
incremental, 892, 938
initialization, 889–890
local maxima, 886, 888–890
log-linear model, 955–956
MAP, 898, 940
noisy-or, 936
overfitting, 891
single family, 937
tree-CPD, 936
variational, 895–897
empirical distribution
Gaussian, 722
endogenous variable, 1027
energy function, 124
canonical, 129
restricted, 592
submodular, 590, 595
truncation, 602
energy functional, 385, 450, 881–882, 905, 914, 940
convex, 416, 962
energy term, 385
entropy term, 385
factored, 386–387, 411
optimization, 411–414
generalized, 414–428
Gibbs distribution, 458
optimization, 459–468
temperature-weighted, 582–585
energy minimization, 553, 599
entanglement, 656–660
entropy, 477, 1138–1142
Bayesian network, 271
conditional, 1139
convex, 417
exponential family, 270
factored, 386, 964
Gaussian, 270
joint, 1139
Markov network, 270
relative, 1141
conditional, 1142
weighted approximate, 415
EP, *see* expectation propagation
equivalent sample size, 740
error
absolute, 290, 544
relative, 291, 491, 544
estimator, 1145
Bayesian, *see* Bayesian estimation
consistent, 769
MAP, *see* MAP estimation
maximum likelihood, *see* maximum likelihood estimation
representation independence, 752–754
unbiased, 1145
variance, 495
event, 15
measurable, 16
evidence, 26
evidence retraction, 339
expectation
linearity of, 32
random variable, 31
expectation maximization, *see* EM
expectation propagation, 430, 441, 444, 664
and belief propagation, 482
convergence point, 447
Gaussian
message passing, 621
mixture, 621–626, 686–688
nonlinear, 630, 637–642
message passing, *see* message passing, expectation propagation
expectation step, *see* E-step
Expectimax, 1087
expert system, 67
expert systems, 13
explaining away, *see* reasoning, intercausal, 55, 196
exponential family, 261, 442, 874, 879
Bayesian network, 268–269
Bernoulli, 265
composition, 266
CPD, 267
EM, 874
factor, 266
Gaussian, 263
invertible, 263, 278, 283
linear, 264, 757
linear Gaussian, 267
multinomial, 265
parameter estimation, 732
exponential time, 1148

factor, 5, 104, 296

division, 365
expected utility, 1108
generalized, 342, 1130
joint, 1103–1107, 1130, 1131
log-space, 360
marginalization, 297, 360, 378
maximization, 555
nonnegative, 104
operations, 358–361
stride, 358
product, 107, 359
reduction, 111, 303
scope, 104
set, 432
marginalization, 432
product, 432
factor graph, 123, 154, 418
factorization, 50
bayesian network, 62
factor graph, 123
Markov network, 109
faithful, 72, 786
faithfulness assumption, 1042
family score, 805
feature
indicator, 125
linear dependence, 132
log-linear model, 125
features, 50
filtering, 652
assumed density, 664
bootstrap, 668
particle, 667–674, 680
collapsed, 674, 693, 694
posterior, 671
Rao-Blackwellized, 674
recursive, 654
state-observation model, 653–654
fixed point
equations, 482
fixed-point, 402
equations, 390, 412, 424, 447, 451, 458, 479
forest, 38
forward pass, 654
forward sampling, 488–492, 541
convergence bounds, 490–491
estimator, 490
sample size, 490, 544
forward-backward algorithm, 337, 655
free energy, 385
Bethe, 414
frequentist interpretation, 16
function
concave, 41
convex, 41

game theory, 1130
Gamma distribution, *see* distribution, Gamma
Gamma function, 735, 798
Gaussian, 28, 1144
Bayesian network, 251–254, 1084
belief propagation, 612–614
clique tree, 611–612
covariance matrix, 247
exponential family, 264
independencies, 250–251, 258
information matrix, 248
linearization, 631–637, 650
incremental, 640
mean vector, 247
mixture, 190, 616
collapsing, 620–621, 624–626, 685–688
pruning, 685
MRF, 254–257
diagonally dominant, 255
pairwise normalizable, 256, 614
walk-summable, 648
multivariate, 247–251
normalizable, 622–624, 639
standard, 28, 248
Gaussian processes, 778
general pseudo-Bayes, 685, 687
generalization, 704–708, 784
generalized linear model, 178
generative training, 709
genetic inheritance, 57–60
GES algorithm, 821
Gibbs distribution, 108
parameterization, *see* Markov network, parameterization
reduced, 111
Gibbs sampling, 505–507, 512–515, 547
block Gibbs, 513
collapsed, 531, 549, 550, 1056
incomplete data, 901–904, 929, 940
continuous state, 644
incomplete data, 899–904
Markov chain, 512

regularity, 514
stationary distribution, 512, 546
goodness of fit, 708, 839
GPB, *see* general pseudo-Bayes
gradient, 1162
ascent, 863, 1163–1166
Bayesian network, 867–868
conjugate, 1166
convergence, 887
L-BFGS, 950, 991
line search, 1164
Bayesian network, 863–866, 936–937
log-likelihood, 864
chain rule, 864
Gaussian, 937
hidden variable, 937
log-linear model, 948
partition function, 947–948
unstable, 962
grafting, 992
graph
acyclic, 37
chordal, 38, 155, 374
connected, 36
directed, 34
moralized
Bayesian network, 134
chain graph, 148
singly connected, 38
skeleton, 77
triangulated, 38
undirected, 34
undirected version, 34
graph cut, 588
ground
Bayesian network, 217, 221, 224
Gibbs distribution, 229
random variable, 215
guard, 223

Hammersley-Clifford theorem, 116, 1077
Hessian, 1163
Bayesian network
incomplete data, 909
log-likelihood, 950
Markov network, 983
partition function, 947
hidden Markov model, 146, 203, 208, 952
coupled, 148, 204
duration, 244
factorial, 204, 482
hierarchical, 210, 244
mixed memory, 244
phylogenetic, 206, 483
segment, 244
hidden variable, 65, 713, 849, 925–932
cardinality, 928–930
model selection, 928
hierarchical, 931
information, 926–928
overlapping, 931
partition, 929
hierarchical Bayes, 765, 779
HMM, *see* hidden Markov model
Hoeffding bound, 490, 771, 1145
holdout testing, 705–708, 795
Hugin, 377
algorithm, *see* message passing, belief update
hybrid network, 186
hyperbolic tangent, 403
hyperparameter, 958
Beta, 735
Dirichlet prior, 738
hierarchical distribution, 765
hypothesis space, 702, 712, 718, 785
hypothesis testing, 787–790
decision rule, 788
deviance, 788
multiple hypotheses, 790, 843
null hypothesis, 787
p-value, 789, 843

I-equivalence, 76, 784, 815
class, 76, 815, 821
I-map, 60
Markov network
construction, 120–122
minimal, 79, 786
construction, 79–81
I-projection, 274, 282, 383
Gaussian, 274
ICI, *see* independence, causal
ICU-Alarm, 749, 796, 802, 820, 830, 885
identifiability, 702, 861
Bayesian network structure, 784, 841
hidden variable, 861
incomplete data, 860–862
intervention query, 1055

local, 862
identity resolution, *see* correspondence, 165, 532
IID, 698, 1144
image denoising, 112
image registration, 532
image segmentation, 113, 478
immorality, 78
potential, 86
importance sampling, 494–505, 545, 547, 966, 1004
adaptive, 542
annealed, 543, 548
backward, 505
Bayesian network, 498–505
collapsed, 527–530
normalized, 496–498, 503, 545
bias, 497
estimator, 497
variance, 497
sample size
effective, 498
sequential, 667–672
variance, 671
unnormalized, 494–496, 502
bias, 495
estimator, 495
variance, 495
incremental update, 369–370
indegree, 34, 804
bounded, 85, 786, 787, 811, 814, 826, 841–842
independence, 23–25
causal, 182, 196, 197
symmetric, 183
conditional
continuous, 31
events, 24
random variables, 24
context-specific, 162, 171–175, 196, 1127
events, 23
marginal, 24
persistent, 657
properties, 24–25, 154
contraction, 25
decomposition, 25
intersection, 25
strong union, 154
symmetry, 24
transitivity, 154
weak union, 25
test, 783, 786–790, 843, 848
independence test
Markov network, 979–981
independencies
Bayesian network, 56–57
global, 72
local, 57
chain graph
global, 151
local, 150
pairwise, 150
distribution, 60
Gaussian, *see* Gaussian, independencies
inclusion, 94
Markov network, 117–120
global, 115
local, 118, 120–122, 979
pairwise, 118, 120–122, 154, 979
indicator function, 32
induced width, 310
inference, 5
inferential loss, 1080
influence diagram, 93, 1089–1090
expected utility, 1093–1094
limited memory, 1093
reduction, 1120, 1132
influence graph, 658
information edge, 1090
irrelevant, 1119–1121
information form, 248
information state, 1091
insurance premium, 1066
interface, 464
interface variable, 202
intervention, 1092, 1112
ideal, 1010
query, 1010, 1015
identifiability, 1017–1026, 1031–1034
simplification, 1018–1026, 1055
Ising model, 126, 127
iterated conditional modes, 599
iterative proportional fitting, 998
iterative proportional scaling, 998, 1002
$I_{\mathcal{X}}$-equivalence, 1049

Jensen inequality, 41
join tree, *see* clique tree
junction tree, *see* clique tree

k-means, 877
K2 prior, 806, 844
Kalman filter, 211, 259, 676–684
  extended, 212, 631, 678
  information form, 677
  observation update, 677
  state transition update, 676
  unscented, 635, 678
kernel density, 730
KL-divergence, *see* entropy, relative
knowledge discovery, 701, 783
knowledge-based model construction, 241, 242, 651

label bias problem, 953
Lagrange multipliers, 388, 868, 1168–1172
language model, 209
Laplace's correction, 735
latent Dirichlet allocation, 769
latent variable, 1012
latent variable network, 1048
Lauritzen's algorithm, 626
Lauritzen-Spiegelhalter algorithm, *see* message passing, belief update
leaf, 38
leak probability, 176
lifted inference, 689
likelihood, 699
  Bayesian network, 723–726
  conditional, 701, 725, 950
  decomposability, 723–726, 857
    global, 725, 859
    local, 726, 859
    shared parameters, 755
  function, 719, 721
  incomplete data, 856–860
    computational complexity, 860
  local, 725
  log-likelihood, 699
  log-linear model, 944–949
    incomplete data, 954–955
likelihood score, 805
likelihood weighting, 493–494, 541
  data dependent, 502
    expected sample size, 502
  DBN, 665–667
  estimator, 493, 500
  normalized, 503, 504
  ratio, 502, 504
likelihood, marginal, *see* marginal likelihood
linear program, 579
  integer, 577
  optimization variables, 577
  relaxation, 576, 579
local consistency polytope, 412, 477, 580, 964
local maximum, 1156
local probability model, 53
log-likelihood, *see* likelihood, 699, 719
  expected, 699, 878–881
log-linear model, 125, 155, 946
  shared parameters, 228, 965, 1002
log-odds, 179
logical variable, 213
logit, *see* sigmoid
loop, 38
loopy belief propagation, *see* belief propagation, loopy
loss function, 699
  0/1 loss, 701
  Hamming loss, 701, 978
  log-loss, 699
lottery, 1059, 1060
  compound, 1062
  preference, 1060
lower bound, 386, 412, 469–473, 897
  variational, *see* variational, lower bound

M-projection, 274, 277–283, 383, 433, 443, 620, 621, 624, 632, 774, 1170–1171
  Bayesian network, 284
  chain network, 280, 284
  exponential family, 278
  factor set, 433–436
  Gaussian, 274, 279, 283
M-step, 873, 874, 907
MAP, *see* query, marginal MAP, 26, 574
  assignment, 534, 537, 1155
  computational complexity, 551–552
  integer program, 577–579
  k-best, 559, 601–603, 977, 1005
  linear program, 579–581
  marginal, 27, 554, 559–561, 595
MAP estimation, 751, 753, 898, 983
  Beta, 754
  log-linear model, 958–961, 984–985
    block L_1 prior, 984
    Gaussian prior, *see* regularization, L_2
    hyperbolic prior, 1003

L1 prior, 988–992
Laplacian prior, *see* regularization, L_1
margin-based estimation, 976–978
marginal independence, *see* independence
marginal likelihood, 738, 744, 795–799, 826
approximation, 909–916
BIC, 911–912, 915
candidate, 913–915
Cheeseman-Stutz, 912–913, 915
Laplace, 909–911, 915
variational, 914
marginal MAP, *see* MAP, marginal, 685
computational complexity, 552, 560–561
marginal polytope, 411, 477, 580
marginalization, *see* factor, marginalization
strong, 627
weak, 621–630
Markov assumption
dynamical system, 201
Markov blanket, 512
Bayesian network, 135, 155
distribution, 121
undirected graph, 118, 980
Markov chain, 507
conductance, 519
ergodic, 510
homogeneous, 507
kernel, 511
mixing, 515, 519–520, 543, 831, 832
empirical, 522–523
multi-kernel, 511, 546
periodic, 510
reducible, 510, 546
regular, 510, 546
reversible, 515
temperature, 524
transition model, 507
Markov chain Monte carlo, *see* MCMC
Markov decision process, 1129
Markov inequality, 40
Markov model, *see* hidden Markov model
Markov network, 5, 103–133
decomposition, 155
pairwise, 110, 404, 478
parameterization, 106–109
canonical, 129–132, 154
redundancy, 132–133, 948
tree, 195
reduced, 111
utility, 1076
Markov random field, *see* Markov network, 105
labeling, 127, 547
metric, 128, 588–595
semimetric, 128, 588–595
max-calibrated, 563, 574
Max-Clique Problem, 1152
max-margin, 1005
max-marginal, 553, 562, 563
decoding, 553, 556–559, 565–567
pseudo, *see* pseudo-max-marginal
ratio optimality, 566
unambiguous, 553
max-product, 552, 582
max-sum, 553, 577, 1117
max-sum-product, 559
maximization step, *see* M-step
maximum entropy, 956–958
approximate, 964–965
distribution, 1169
expectation constraints, 956
maximum entropy Markov model, 952
maximum expected utility, *see* MEU
maximum likelihood estimation, 719, 722
Bayesian network, 723–732
conditional random field, 950–953
consistent, 949, 1002
Gaussian, 722, 778
incomplete data
computational complexity, 887
linear Gaussian, 728–730
log-linear model, 949–950
dual, 956–958, 1002
using belief propagation, 963–965
using MCMC, 966–967
multinomial, 722
plate models, 757–760
shared parameters, 756–761
table-CPD, 725
maximum spanning tree, 374
MCMC, *see* Markov chain, 507, 644, 673, 966, 975, 1159
burn-in time, 519
collapsed, 531–532, 831
estimator, 521
variance, 521–522
Gibbs sampling, *see* Gibbs sampling
Metropolis-Hastings, *see* Metropolis-Hastings
network structures, 829–831

reversible jump, 935
sampling, 508, 520–523
autocovariance, 521, 522
variable ordering, 831–832
mean field, 449–456, 895, 906
algorithm, 454–456
cluster, 467
convergence point, 451–453
energy, 449–450
mean prediction, 740
medical diagnosis, 51, 67–68, 177, 183, 197, 1124
message decoding, 393
turbocode, 395
message passing
belief-update, 364–368
CLG network, 624–626
max-product, 563
clique tree, 351–352
DBN, 654–655
expectation propagation
belief-update, 440–442
exponential family, 442–445
Gaussian, 641
sum-product, 437–439
max-product, 563, 603
counting numbers, 573
order-constrained, 623, 639
region graph, 425–428, 480–481
sum-product, 352, 368, 397, 413
generalized, 418
sum-product-divide, 365–368
meta-network, 742, 858, 899
global decomposition, 743
local decomposition, 746
shared parameters, 763
metric, 127
Metropolis-Hastings, 516–518, 542, 547, 832, 942, 1159
acceptance probability, 517
collapsed
incomplete data, 940
continuous state, 644
random walk, 645, 903
MEU
principle, 1061
strategy
decision rule, 1107, 1108
decision tree, 1098–1100
influence diagram, 1094, 1115, 1131
value, 1094, 1119
micromort, 1070, 1081
min-fill, 314
weighted, 314
min-neighbors, 314
min-weight, 314
minimax risk, 1083
minimum description length, 802
missing at random, 854, 936
missing completely at random, 853
MLE, *see* maximum likelihood estimation
model dimension, 801, 983
model selection, 785, 978
module network, 846
moment matching, 278, 949
Monte Carlo localization, *see* filter, particle, *see* robot, localization, 680, 691
moral graph, 135
MPE, *see* query, MAP
MRF, *see* Markov random field
multiconditional training, 1004
multinet, 170, 195
mutilated network, 499, 1014
interventional, 1014–1017, 1044
proposal distribution, 499–500, 530
mutual information, 789, 792, 848, 1140
chain rule, 41
conditional, 41

naive Bayes, 49, 727
Bernoulli, 767
clustering, 875, 877, 915
multinomial, 767
tree augmented, 842
naive Markov, 144, 197, 710
natural bounds, 1033
negative definite, 1163
neighbor, 34
network polynomial, 304, 339, 378
noise parameter, 176
noisy-or model, *see* CPD, noisy-or
normal-Gamma distribution, 780
NP-hardness, 1150–1153
CSI-separation, 196
elimination ordering, 310
inference
approximate, 291–292
exact, 288–290
MAP, 551

polytree CLG, 617
reduction, 1152
structure learning
directed, 811, 841
undirected, 1000
triangulation, 313
numerical integration, 633
exact monomials, 634–637
precision, 635
Gaussian quadrature, 633–634
precision, 633
integration rule, 633, 634
precision, 633

object, 213
object skeleton, 214, 229
object uncertainty, 233
object-oriented Bayesian networks, 192
objective function, 702, 718, 1154
concave
over the constraints, 417
observability
model, 851
variable, 851
observation model, 207
observed variable, 24, 71, 114, 142
optimization
constrained, 381, 1167–1171
optimization problem, 1154
outcome, 1061, 1090
anchor, 1064
atomic, 22
space, 15
canonical, 22
overfitting, 704–708, 726, 769, 794, 801, 886

P-map, *see* perfect map
PAC-bound, 709, 770
Bayesian network, 773–776
log-linear model, 991, 1000–1001
multinomial, 771–773
parameter
sharing, *see* shared parameters
space, 720
parameter distribution
Bernoulli, *see* Beta distribution
conjugate prior, 739
Gaussian, *see* normal Gamma distribution
multinomial, *see* Dirichlet distribution
prior, 733
parameter independence, 799, 834, 857
global, 742, 805–806, 837
local, 747
parameter modularity, 805
CPD-tree, 835
parameter posterior, 734, 738
parameter prior, *see* parameter distribution, prior, 738
Bayesian network, 748, 805–806
conjugate, 737
log-linear model
conjugate, 961
L_1, 959
L_2, 958
parameters, 46, 720
independent, 46, 259, 801
incomplete data, 912
legal, 262
natural, 263
function, 262
space, 264
space, 262
parametric family, 261, 720
parametric model, 720
parent, 34
partial ancestral graph, 1049–1051
partial correlation coefficient, 259
partially directed acyclic graph, *see* PDAG, 148
particle, 487
collapsed, 487, 526, 543, 674
data completion, 903–904, 940
parameter, 901–903
deterministic, 536–540, 675
deterministic search, 549
weighted, 493
particle filtering
smoothing, 692
partition function, 105, 108, 262, 543
approximate, 966
convex, 947
lower bound, 386, 470
upper bound, 1004–1005
Pascal's wager, 1082
path, 36
active, 114
Pathfinder, 67
PDAG, 37, 843
boundary, 34, 149

class, 87, 786, 821, 1042
PDF, *see* probability density function
peeling, 337
perfect map, 81, 787
construction, 83–92
persistence
edge, 204, 658
variable, 204
piecewise training, 1003, 1004
plate, 217
intersection, 218
nested, 218
plate model, 216–222, 837
text, 767
plateau, 1156
point estimate, 737
polynomial time, 1148
polytree, 38, 313, 340, 552, 617
positive definite, 248
positive semi-definite, 248
posterior, 26
potential
edge, 110
node, 110
Potts model, 127
prediction, 652
preference independence, 1071–1072
prenatal diagnosis, 1079, 1094
prequential analysis, 796
prior, 19
improper, 740
probabilistic context-free grammar, 243
probabilistic finite-state automaton, 209
probabilistic relational model, 223, 837
parameter estimation, 781
probability distribution, *see* distribution
probability query, 26, 287
approximate, 290–292
computational complexity, 288–292
lower bound, 537
reasoning, 54–55
causal, 54
evidential, 55
intercausal, 55
probability theory, 2
probably approximately correct, 709
projection, *see* M-projection; I-projection
proposal distribution, 644, 1160
importance sampling, 494, 498, 528–530, 542
MCMC, *see* Metropolis Hastings, 516
protein structure prediction, 968–969
pseudo-counts, 740
pseudo-marginal, 412, 580
pseudo-max-marginal, 562, 568
decoding, 568–572
pseudo-moment matching, 963, 1004
pseudolikelihood, 970–974
consistent, 972
generalized, 973

QALY, 1070
quadratic program, 976
qualitative probabilistic networks, 94
query variable, 26

random variable, 3, 20–23
Rao-Blackwellization, *see* collapsed sampling
rationality
human, 1067–1068
postulates, 1062–1064, 1084
recall
edge, 1092
imperfect, 1093, 1109, 1116, 1119
perfect, 1092, 1098, 1131
record matching, *see* correspondence
redundant
feature, 133
parameterization, 263
reference class, 17
region graph, 419–428, 572
belief propagation, *see* belief propagtion, region graph
calibrated, 421
construction, 421–423
saturated, 422
regularization, 705, 751
block-L_1, 984
L_1, 959, 984
L_2, 958, 984
log-linear model, 958–961
rejection sampling, 491, 643
relation, 213
relational Markov network, 229, 1002
relational skeleton, 224
relational uncertainty, 225, 233
relative entropy, 771
Bayesian network, 273
exponential family, 272

relevance graph, 1114–1115, 1131
renormalization, 287, 339
reparameterization, 574, 868
  max-product
    clique tree, 564–565
    cluster graph, 568
    counting numbers, 574
  sum-product
    clique tree, 362
    cluster graph, 399
response variable, 1028–1030, 1035
  constraints, 1031–1033
risk, 700, 1066
  averse, 1066
  empirical, 700
  excess, 709, 774
  neutral, 1067
  seeking, 1067
RoboSoccer, 1117
robot
  localization, 187, 678–684
  mapping, 681, 892–893, 938
  SLAM, 681, 694
$\mathcal{RP}$, 1153
rule, 166
  product, 330
  reduced, 172
  scope, 166
  split, 331
  sum, 330
running intersection property, *see* clique tree, running intersection property

s-reachable, 1112–1114, 1131
Saint Petersburg paradox, 1065
sample complexity, 709
sample size, 501
search, 675
  assignment, 536–540, 595–597
  beam, 595, 675, 685, 693, 1158
  branch-and-bound, 595, 603, 604, 1160–1161
  hill-climbing
    first-ascent, 815, 1155
    greedy, 1155
  local, 595, 812, 814, 985, 1154–1160
  operators, 596, 1154
  random restart, 1159
  randomization, 1158–1160
  space, 595, 812, 1154
  state, 1154
  systematic, 595
  tabu, 596, 816, 1156
selection bias, 1013
selector variable, 165
semimetric, 128
sensitivity analysis, 67, 95, 305, 339
separation, 115
  completeness, 116–117
    CA-independence, 1077
  soundness, 115–116
sepset, *see* clique tree, sepset, *see* cluster graph, sepset
sequence labeling, 952
shared parameters, 754, 780–781
  global, 755–760
  local, 760–761
$\#\mathcal{P}$, 1153
shrinkage, 243, 764
sigmoid, 145, 179
similarity network, 95, 171
Simpson's paradox, 1015–1016, 1021
simulated annealing, 524, 1159
smoothing, 652
  computational complexity, 692
  particle, 692
spanning forest, *see* spanning tree
spanning tree
  maximum weight, 809, 1146, 1148
speech recognition, 209, 675
standard deviation, 33
standard gamble, 1069
state-observation model, 207
stationary distribution, 509–511
stationary point, 1162
stereo reconstruction, 113, 593
stick-breaking prior, 930
Stirling's approximation, 843
strategic relevance, 1110–1115
strategy, 1087
  complete, 1091
  MEU, *see* MEU strategy
structural uncertainty, 232
structure discovery, 825
  confidence estimation, 825
  network features, 825, 827–828
structure learning
  constraint-based, 785–790, 1042
    Markov network, 979–981, 1005

score-based, 785, 790–824
undirected model, 981–995
convergence, 990
global maximum, 989
hypothesis space, 981–982
L_1 prior, 988–992
structure modularity, 804
structure score
Bayesian, 794–807, 843, 983–984
decomposable, 799–801, 805
BIC, 802, 843, 911, 983
consistent, 803, 822
decomposability, 808, 818–820, 917–919, 986
decomposable, 805
equivalence, 808
Laplace approximation, 983
likelihood, 791–794, 982–983
decomposable, 792
MAP, 984–985
L_1, 984, 988–995
score equivalence, 807, 821, 844
structure prior, 804
tree-CPD, 834
template model
decomposable, 837
tree-CPD
decomposable, 834
structure search, 807–824, 1155
computational complexity, 809, 811, 814–815, 818–820
delta score, 818, 917
hidden variable
initialization, 932
I-equivalence classes, 821–824
incomplete data, 917–925
heuristics, 919–920
structural EM, 920–925, 932, 941
local maximum, 815–818
operators, 812–814, 845
edge addition, 812
edge deletion, 812
edge reversal, 812, 813
reinsertion, 847
ordering space, 848
parent constraints, 845
plateau, 815
template model, 837
tree-CPD, 835–836
delta-score, 846
operators, 835
trees, 808–809
undirected model, 985–995
computational complexity, 987
delta-score, 987, 992–995
gain heuristic, 993–995, 1005
gradient heuristic, 992
local maximum, 988
variable ordering, 809–811
structured variational, 448–469, 895
algorithm, 459–468, 482
convergence point, 458, 482
update, 460–468, 482
subgraph
complete, 35
induced, 35
subjective interpretation, 17
subutility function, 1071, 1073, 1117
sufficient statistics, 721
aggregate, 756
Bernoulli, 265
collection, 819–820
expected, 278, 871–874, 880
belief propagation, 962
conditional random field, 951
log-linear model, 949
MAP assignment, 967–968
MCMC, 966–967, 1004
function, 262
Gaussian, 263, 721
interventional data, 1044–1046, 1056
log-linear model, 947
multinomial, 265, 721
sum-max-sum rule, 1098
sum-product, 299, 582, 611
message passing, *see* message passing, sum-product
support vector machine, 999
survey propagation, 601
Swendson-Wang algorithm, 547
system state, 200

t-node, 1085
table-CPD, *see* CPD, table
target distribution, 494
target tracking, 678–684
target variable, 142
Taylor series, 631
temperature, 582

temperature parameter, 126, 524, 1160
template
  variable
    instantiated, *see* ground random variable
template model
  dependency graph, 227, 245
  factor, 203, 216
    instantiated, 216
  feature, 228
  lifted inference, 689
  parent, 221, 223
  structure learning, 837–838, 846
  variable, 200, 213
template variable, *see* attribute
temporal ordering, 1092, 1097, 1131
test set, 705
time slice, 201
time trade-off, 1069
topological ordering, 36, 62, 1146
trail, 36
  active, 71
    minimal, 100
training set, 705, 720
trajectory, 200
transition model
  dynamical system, 202
  state-observation model, 207
tree, 38, 808
tree reparameterization, *see* belief propagation, tree reparameterization
tree-CPD, *see* CPD, tree-CPD, 834, 936
  structure learning, 834–836, 845
tree-width, 310
  bounded, 982, 1000
triangle inequality, 1140
triangulation, 139, 313, 374
troubleshooting, 166, 1027, 1037, 1055, 1124, 1132
truncated norm, 128, 603
TRW, *see* belief propagation, tree-reweighted

uncertainty, 2
unrolled Bayesian network, 204
unscented transformation, 634
upward closure, 35, 136
utility, 1060
  additive independence, 1074–1075
  CA-independence, 1075–1078, 1084
  expected, 1060, 1061, 1064, 1087, 1093
  GA-independence, 1078–1079, 1084
  independence, 1072–1073, 1081
  variable, 1090
utility function, 1061
  curve, 1065–1067
  decomposition, 1073
    additive, 1073–1080, 1117
    multilinear, 1073
    multiplicative, 1073
  distribution, 1084
  elicitation, 1069, 1080–1081
  factorization, 1076
  human life, 1069–1070
  indifference point, 1069
  money, 1065–1066

v-structure, 71
validation set, 708, 891
value of control, 1132
value of information, 1121–1125, 1132
  myopic, 1125, 1126
  perfect, 1122
variable elimination, 299, 372, 1099
  and conditioning, 319–322, 340
  causal independence, 325–329
  chordal graph, 310–313
  cliques, 308
  computational complexity, 305–310, 336
  context-specific independence, 329–334
  expected utility, 1100–1107
  factor semantics, 301, 338
  generalized, 342–343, 1103–1107, 1130, 1131
  induced graph, 306–310
  max-product, 556
    traceback, 558
  max-sum-product, 559–561
    traceback, 561, 601
  ordering, 299–301, 310
    computational complexity, 310
    constrained, 561, 596, 629, 1100, 1109
    heuristics, 310–315, 340
    maximum cardinality, 312, 340
  rule-based CPDs, 329–334, 341
  sum-product, 299
  variational, 470–473, 483
  with evidence, 303
variable ordering, 79–81, 809, 826
variance, 33
variational
  Bayesian network, 483

lower bound, 469–472, 484
method, 386, 469–473
mixture distribution, 484
parameter, 470
sigmoid, 483
variable elimination, 470, 472–473
variational Bayes, 904–908
variational distance, 1143
variational, Markov network, *see* Gibbs variational
visual-analog scale, 1069
Viterbi algorithm, 598, 675
Viterbi training, 967

witness, 85