An Introduction to Deterministic Annealing

Fabrice Rossi

SAMM – Université Paris 1

2012
Optimization

- in machine learning:
  - optimization is one of the central tools
  - methodology:
    - choose a model with some adjustable parameters
    - choose a goodness of fit measure of the model to some data
    - tune the parameters in order to maximize the goodness of fit
  - examples: artificial neural networks, support vector machines, etc.

- in other fields:
  - operational research: project planning, routing, scheduling, etc.
  - design: antenna, wings, engines, etc.
  - etc.
Easy or Hard?

- is optimization computationally difficult?
Easy or Hard?

- is optimization computationally difficult?
- the convex case is relatively easy:
  - $\min_{x \in C} J(x)$ with $C$ convex and $J$ convex
  - polynomial algorithms (in general)
  - why?
Easy or Hard?

- is optimization computationally difficult?
- the convex case is relatively easy:
  - $\min_{x \in C} J(x)$ with $C$ convex and $J$ convex
  - polynomial algorithms (in general)
  - why?
    - a local minimum is global
    - the local tangent hyperplane is a global lower bound
Easy or Hard?

- is optimization computationally difficult?
- the convex case is relatively easy:
  - $\min_{x \in C} J(x)$ with $C$ convex and $J$ convex
  - polynomial algorithms (in general)
  - why?
    - a local minimum is global
    - the local tangent hyperplane is a global lower bound
- the non convex case is hard:
  - multiple minima
  - no local to global inference
  - NP hard in some cases
In machine learning

- convex case:
  - linear models with(out) regularization (ridge, lasso)
  - kernel machines (SVM and friends)
  - nonlinear projections (e.g., semi-define embedding)

- non convex:
  - artificial neural networks (such as multilayer perceptrons)
  - vector quantization (a.k.a. prototype based clustering)
  - general clustering
  - optimization with respect to meta parameters:
    - kernel parameters
    - discrete parameters, e.g., feature selection
In machine learning

▶ convex case:
  ▶ linear models with(out) regularization (ridge, lasso)
  ▶ kernel machines (SVM and friends)
  ▶ nonlinear projections (e.g., semi-define embedding)

▶ non convex:
  ▶ artificial neural networks (such as multilayer perceptrons)
  ▶ vector quantization (a.k.a. prototype based clustering)
  ▶ general clustering
  ▶ optimization with respect to meta parameters:
    ▶ kernel parameters
    ▶ discrete parameters, e.g., feature selection

▶ in this lecture, non convex problems:
  ▶ combinatorial optimization
  ▶ mixed optimization
Particular cases

- Pure combinatorial optimization problems:
  - a solution space $\mathcal{M}$: finite but (very) large
  - an error measure $E$ from $\mathcal{M}$ to $\mathbb{R}$
  - goal: solve
    \[ M^* = \arg \min_{M \in \mathcal{M}} E(M) \]
  - example: graph clustering

- Mixed problems:
  - a solution space $\mathcal{M} \times \mathbb{R}^p$
  - an error measure $E$ from $\mathcal{M} \times \mathbb{R}^p$ to $\mathbb{R}$
  - goal: solve
    \[ (M^*, W^*) = \arg \min_{M \in \mathcal{M}, W \in \mathbb{R}^p} E(M, W) \]
  - example: clustering in $\mathbb{R}^n$
Particular cases

- Pure combinatorial optimization problems:
  - a solution space $\mathcal{M}$: finite but (very) large
  - an error measure $E$ from $\mathcal{M}$ to $\mathbb{R}$
  - goal: solve
    $$M^* = \arg \min_{M \in \mathcal{M}} E(M)$$
  - example: graph clustering

- Mixed problems:
  - a solution space $\mathcal{M} \times \mathbb{R}^p$
  - an error measure $E$ from $\mathcal{M} \times \mathbb{R}^p$ to $\mathbb{R}$
  - goal: solve
    $$(M^*, W^*) = \arg \min_{M \in \mathcal{M}, W \in \mathbb{R}^p} E(M, W)$$
  - example: clustering in $\mathbb{R}^n$
Graph clustering

Goal: find an optimal clustering of a graph with $N$ nodes in $K$ clusters

- $\mathcal{M}$: set of all partitions of $\{1, \ldots, N\}$ in $K$ classes (the asymptotic behavior of $|\mathcal{M}|$ is roughly $K^{N-1}$ for a fixed $K$ with $N \to \infty$)

- assignment matrix: a partition in $\mathcal{M}$ is described by $N \times K$ matrix $M$ such that $M_{ik} \in \{0, 1\}$ and $\sum_{k=1}^{K} M_{ik} = 1$

- many error measures are available:
  - Graph cut measures (node normalized, edge normalized, etc.)
  - Modularity
  - etc.
Graph clustering

- original graph
Graph clustering

- original graph
- four clusters
Graph visualization

- 2386 persons: unreadable

- woman
- bisexual man
- heterosexual man

Maximal modularity clustering in 39 clusters
Graph visualization

- 2386 persons: unreadable
- maximal modularity *clustering* in 39 clusters
Graph visualization

- 2386 persons: unreadable
- maximal modularity *clustering* in 39 clusters
Graph visualization

- 2386 persons: unreadable
- maximal modularity clustering in 39 clusters
- hierarchical display
Vector quantization

- $N$ observations $(x_i)_{1 \leq i \leq N}$ in $\mathbb{R}^n$
- $\mathcal{M}$: set of all partitions
- quantization error:

$$E(M, W) = \sum_{i=1}^{N} \sum_{k=1}^{K} M_{ik} \| x_i - w_k \|^2$$

- the “continuous” parameters are the prototypes $w_k \in \mathbb{R}^n$
- the assignment matrix notation is equivalent to the standard formulation

$$E(M, W) = \sum_{i=1}^{N} \| x_i - w_{k(i)} \|^2,$$

where $k(i)$ is the index of the cluster to which $x_i$ is assigned
Example
Example
Combinatorial optimization

- a research field by itself
- many problems are NP hard...
Combinatorial optimization

- a research field by itself
- many problems are NP hard... and one therefore relies on heuristics or specialized approaches:
  - relaxation methods
  - branch and bound methods
  - stochastic approaches:
    - simulated annealing
    - genetic algorithms
Combinatorial optimization

- a research field by itself
- many problems are NP hard... and one therefore relies on heuristics or specialized approaches:
  - relaxation methods
  - branch and bound methods
  - stochastic approaches:
    - simulated annealing
    - genetic algorithms
- in this presentation: deterministic annealing
Outline

Introduction

Mixed problems
  Soft minimum
  Computing the soft minimum
  Evolution of $\beta$

Deterministic Annealing
  Annealing
  Maximum entropy
  Phase transitions
  Mass constrained deterministic annealing

Combinatorial problems
  Expectation approximations
  Mean field annealing
  In practice
Outline

Introduction

Mixed problems
  Soft minimum
  Computing the soft minimum
  Evolution of $\beta$

Deterministic Annealing
  Annealing
  Maximum entropy
  Phase transitions
  Mass constrained deterministic annealing

Combinatorial problems
  Expectation approximations
  Mean field annealing
  In practice
Optimization strategies

- mixed problem transformation

\[ (M^*, W^*) = \arg \min_{M \in \mathcal{M}, W \in \mathbb{R}^p} E(M, W), \]

- remove the continuous part

\[
\min_{M \in \mathcal{M}, W \in \mathbb{R}^p} E(M, W) = \min_{M \in \mathcal{M}} \left( M \mapsto \min_{W \in \mathbb{R}^p} E(M, W) \right)
\]

- or the combinatorial part

\[
\min_{M \in \mathcal{M}, W \in \mathbb{R}^p} E(M, W) = \min_{W \in \mathbb{R}^p} \left( W \mapsto \min_{M \in \mathcal{M}} E(M, W) \right)
\]

this is not alternate optimization
Optimization strategies

- mixed problem transformation

\[
(M^*, W^*) = \arg \min_{M \in \mathcal{M}, W \in \mathbb{R}^p} E(M, W),
\]

- remove the continuous part

\[
\min_{M \in \mathcal{M}, W \in \mathbb{R}^p} E(M, W) = \min_{M \in \mathcal{M}} \left( M \mapsto \min_{W \in \mathbb{R}^p} E(M, W) \right)
\]

- or the combinatorial part

\[
\min_{M \in \mathcal{M}, W \in \mathbb{R}^p} E(M, W) = \min_{W \in \mathbb{R}^p} \left( W \mapsto \min_{M \in \mathcal{M}} E(M, W) \right)
\]

- this is not alternate optimization
Alternate optimization

- elementary heuristics:
  1. start with a random configuration $M^0 \in \mathcal{M}$
  2. compute $W^i = \arg\min_{W \in \mathbb{R}^p} E(M^{i-1}, W)$
  3. compute $M^i = \arg\min_{M \in \mathcal{M}} E(M, W^i)$
  4. go back to 2 until convergence

- e.g., the k-means algorithm for vector quantization:
  1. start with a random partition $M^0 \in \mathcal{M}$
  2. compute the optimal prototypes with $W^i_k = \frac{1}{\sum_{j=1}^{N} M^i_{j-1} \delta_{jk}} \sum_{j=1}^{N} M^i_{j-1} \delta_{jk} x_j$
  3. compute the optimal partition with $M^i_{jk} = 1$ if and only if $k = \arg\min_{1 \leq l \leq K} \|x_j - W^i_l\|_2$
  4. go back to 2 until convergence

alternate optimization converges to a local minimum
Alternate optimization

- elementary heuristics:
  1. start with a random configuration $M^0 \in \mathcal{M}$
  2. compute $W^i = \arg \min_{W \in \mathbb{R}^p} E(M^{i-1}, W)$
  3. compute $M^i = \arg \min_{M \in \mathcal{M}} E(M, W^i)$
  4. go back to 2 until convergence

- e.g., the k-means algorithm for vector quantization:
  1. start with a random partition $M^0 \in \mathcal{M}$
  2. compute the optimal prototypes with
     \[
     W^i_k = \frac{1}{\sum_{j=1}^N M^{i-1}_{jk}} \sum_{j=1}^N M^{i-1}_{jk} x_j
     \]
  3. compute the optimal partition with $M^i_{jk} = 1$ if and only if
     \[k = \arg \min_{1 \leq l \leq K} \|x_j - W^i_l\|^2\]
  4. go back to 2 until convergence
Alternate optimization

- elementary heuristics:
  1. start with a random configuration $M^0 \in \mathcal{M}$
  2. compute $W^i = \arg \min_{W \in \mathbb{R}^p} E(M^{i-1}, W)$
  3. compute $M^i = \arg \min_{M \in \mathcal{M}} E(M, W^i)$
  4. go back to 2 until convergence

- e.g., the $k$-means algorithm for vector quantization:
  1. start with a random partition $M^0 \in \mathcal{M}$
  2. compute the optimal prototypes with
     $\quad W^i_k = \frac{1}{\sum_{j=1}^N M^{i-1}_{jk} \sum_{j=1}^N M^{i-1}_{jk} x_j}$
  3. compute the optimal partition with $M^i_{jk} = 1$ if and only if
     $\quad k = \arg \min_{1 \leq l \leq K} \|x_j - W^i_l\|^2$
  4. go back to 2 until convergence

- alternate optimization converges to a local minimum
Combinatorial first

- let consider the combinatorial first approach:

\[
\min_{M \in \mathcal{M}, W \in \mathbb{R}^p} E(M, W) = \min_{W \in \mathbb{R}^p} \left( W \mapsto \min_{M \in \mathcal{M}} E(M, W) \right)
\]

- not attractive \textit{a priori}, as \( W \mapsto \min_{M \in \mathcal{M}} E(M, W) \):
  - has no reason to be convex
  - has no reason to be \( C^1 \)
Combinatorial first

- let consider the combinatorial first approach:

\[
\min_{M \in \mathcal{M}, W \in \mathbb{R}^p} E(M, W) = \min_{W \in \mathbb{R}^p} \left( W \mapsto \min_{M \in \mathcal{M}} E(M, W) \right)
\]

- not attractive \textit{a priori}, as \( W \mapsto \min_{M \in \mathcal{M}} E(M, W) \):
  - has no reason to be convex
  - has no reason to be \( C^1 \)

- vector quantization example:

\[
F(W) = \sum_{i=1}^{N} \min_{1 \leq k \leq K} \| x_i - w_k \|^2
\]

is neither convex nor \( C^1 \)
Example

Clustering in 2 clusters elements from $\mathbb{R}$
Example

\[ \ln F(W) \]

multiple local minima and singularities: minimizing \( F(W) \) is difficult
Soft minimum

- A soft minimum approximation of $\min_{M \in \mathcal{M}} E(M, W)\]

$$F_\beta(W) = -\frac{1}{\beta} \ln \sum_{M \in \mathcal{M}} \exp(-\beta E(M, W))$$

Solves the regularity issue: if $E(M, .)$ is $C^1$, then $F_\beta(W)$ is also $C^1$
Soft minimum

- a soft minimum approximation of $\min_{M \in \mathcal{M}} E(M, W)$

$$F_\beta(W) = -\frac{1}{\beta} \ln \sum_{M \in \mathcal{M}} \exp(-\beta E(M, W))$$

solves the regularity issue: if $E(M, .)$ is $C^1$, then $F_\beta(W)$ is also $C^1$

- if $M^*(W)$ is such that for all $M \neq M^*(W)$, $E(M, W) > E(M^*(W), W)$, then

$$\lim_{\beta \to \infty} F_\beta(W) = E(M^*(W), W) = \min_{M \in \mathcal{M}} E(M, W)$$
Example

$$E(M, W)$$
Example

\[ \beta = 10^{-2} \]

\[
\frac{\exp(-\beta E(M, W))}{\exp(-\beta E(M^*, W))}
\]
Example

\begin{equation}
\beta = 10^{-1.5}
\end{equation}

\[
\frac{\exp(-\beta E(M, W))}{\exp(-\beta E(M^*, W))}
\]
Example

\[ \beta = 10^{-1} \]

\[ \frac{\exp(-\beta E(M, W))}{\exp(-\beta E(M^*, W))} \]
\[ \beta = 10^{-0.5} \]

\[ \frac{\exp(-\beta E(M, W))}{\exp(-\beta E(M^*, W))} \]
$\beta = 10^0$

$$\frac{\exp(-\beta E(M, W))}{\exp(-\beta E(M^*, W))}$$
\beta = 10^{0.5}

\exp(-\beta E(M,W))

\exp(-\beta E(M^*,W))
Example

$$\beta = 10^1$$

$$\frac{\exp(-\beta E(M, W))}{\exp(-\beta E(M^*, W))}$$
Example

\[ \beta = 10^{1.5} \]

\[
\frac{\exp(-\beta E(M, W))}{\exp(-\beta E(M^*, W))}
\]
Example

\[ \beta = 10^2 \]

\[ \frac{\exp(-\beta E(M, W))}{\exp(-\beta E(M^*, W))} \]
Example
Discussion

- positive aspects:
  - if $E(M, W)$ is $C^1$ with respect to $W$ for all $M$, then $F_\beta$ is $C^1$
  - $\lim_{\beta \to \infty} F_\beta(W) = \min_{M \in \mathcal{M}} E(M, W)$
  - $\forall \beta > 0,$
    
    $$-\frac{1}{\beta} \ln |\mathcal{M}| \leq F_\beta(W) - \min_{M \in \mathcal{M}} E(M, W) \leq 0$$
    
  - when $\beta$ is close to 0, $F_\beta$ is generally easy to minimize
Discussion

- **positive aspects:**
  - if $E(M, W)$ is $C^1$ with respect to $W$ for all $M$, then $F_\beta$ is $C^1$
  - $\lim_{\beta \to \infty} F_\beta(W) = \min_{M \in \mathcal{M}} E(M, W)$
  - $\forall \beta > 0,$
    \[-\frac{1}{\beta} \ln |\mathcal{M}| \leq F_\beta(W) - \min_{M \in \mathcal{M}} E(M, W) \leq 0\]
  - when $\beta$ is close to 0, $F_\beta$ is generally easy to minimize

- **negative aspects:**
  - how to compute $F_\beta(W)$ efficiently?
  - how to solve $\min_{W \in \mathbb{R}^p} F_\beta(W)$?
  - why would that work in practice?
Discussion

- **positive aspects:**
  - if $E(M, W)$ is $C^1$ with respect to $W$ for all $M$, then $F_\beta$ is $C^1$
  - $\lim_{\beta \to \infty} F_\beta(W) = \min_{M \in \mathcal{M}} E(M, W)$
  - $\forall \beta > 0,$
    \[
      - \frac{1}{\beta} \ln |\mathcal{M}| \leq F_\beta(W) - \min_{M \in \mathcal{M}} E(M, W) \leq 0
    \]
  - when $\beta$ is close to 0, $F_\beta$ is generally easy to minimize

- **negative aspects:**
  - how to compute $F_\beta(W)$ efficiently?
  - how to solve $\min_{W \in \mathbb{R}^p} F_\beta(W)$?
  - why would that work in practice?

- **philosophy:** where does $F_\beta(W)$ come from?
Outline

Introduction

Mixed problems
  Soft minimum
  Computing the soft minimum
  Evolution of $\beta$

Deterministic Annealing
  Annealing
  Maximum entropy
  Phase transitions
  Mass constrained deterministic annealing

Combinatorial problems
  Expectation approximations
  Mean field annealing
  In practice
Computing $F_\beta(W)$

- in general computing $F_\beta(W)$ is intractable: exhaustive calculation of $E(M, W)$ on the whole set $\mathcal{M}$
Computing $F_\beta(W)$

- In general, computing $F_\beta(W)$ is intractable: exhaustive calculation of $E(M, W)$ on the whole set $\mathcal{M}$.
- A particular case: clustering with an additive cost, i.e.

$$E(M, W) = \sum_{i=1}^{N} \sum_{k=1}^{K} M_{ik} e_{ik}(W),$$

where $M$ is an assignment matrix (and e.g., $e_{ik}(W) = \|x_i - w_k\|^2$).

- Then

$$F_\beta(W) = -\frac{1}{\beta} \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \exp(-\beta e_{ik}(W))$$

computational cost $O(NK)$ (compared to $\sim K^{N-1}$).
Proof sketch

- let $Z_\beta(W)$ be the **partition function** given by
  \[
  Z_\beta(W) = \sum_{M \in \mathcal{M}} \exp(-\beta E(M, W))
  \]

- assignments in $\mathcal{M}$ are independent and the sum can be rewritten
  \[
  Z_\beta(W) = \sum_{M_1 \in C_K} \ldots \sum_{M_N \in C_K} \exp(-\beta E(M, W)),
  \]

  with $C_K = \{(1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)\}$

- then
  \[
  Z_\beta(W) = \prod_{i=1}^N \sum_{M_i \in C_K} \exp(-\beta \sum_k M_{ik} e_{ik}(W))
  \]
Independence

- additive cost corresponds to some form of conditional independence
- given the prototypes, observations are assigned independently to their optimal clusters
- in other words: the global optimal assignment is the concatenation of the individual optimal assignment
- this is what makes alternate optimization tractable in the K-means despite its combinatorial aspect:

$$\min_{M \in \mathcal{M}} \sum_{i=1}^{N} \sum_{k=1}^{K} M_{ik} \| x_i - w_k \|^2$$
Minimizing $F_\beta(W)$

- assume $E$ to be $C^1$ with respect to $W$, then

$$\nabla F_\beta(W) = \frac{\sum_{M \in \mathcal{M}} \exp(-\beta E(M, W)) \nabla_W E(M, W)}{\sum_{M \in \mathcal{M}} \exp(-\beta E(M, W))}$$

- at a minimum $\nabla F_\beta(W) = 0 \Rightarrow$ solve the equation or use gradient descent

- same calculation problem as for $F_\beta(W)$: in general, an exhaustive scan of $\mathcal{M}$ is needed
Minimizing $F_\beta(W)$

- assume $E$ to be $C^1$ with respect to $W$, then
  
  $$\nabla F_\beta(W) = \frac{\sum_{M \in \mathcal{M}} \exp(-\beta E(M, W)) \nabla W E(M, W)}{\sum_{M \in \mathcal{M}} \exp(-\beta E(M, W))}$$

- at a minimum $\nabla F_\beta(W) = 0 \Rightarrow$ solve the equation or use gradient descent

- same calculation problem as for $F_\beta(W)$: in general, an exhaustive scan of $\mathcal{M}$ is needed

- if $E$ is additive
  
  $$\nabla F_\beta(W) = \sum_{i=1}^{N} \frac{\sum_{k=1}^{K} \exp(-\beta e_{ik}(W)) \nabla W e_{ik}(W)}{\sum_{k=1}^{K} \exp(-\beta e_{ik}(W))}$$
Fixed point scheme

- a simple strategy to solve $\nabla F_\beta(W) = 0$
- starting from a random value of $W$:
  1. compute

$$\mu_{ik} = \frac{\exp(-\beta e_{ik}(W))}{\sum_{l=1}^{K} \exp(-\beta e_{il}(W))}$$

2. keeping the $\mu_{ik}$ constant, solve for $W$

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \mu_{ik} \nabla W e_{ik}(W) = 0$$

3. loop to 1 until convergence
Fixed point scheme

- a simple strategy to solve $\nabla F_\beta(W) = 0$
- starting from a random value of $W$:
  1. compute (expectation phase)

$$
\mu_{ik} = \frac{\exp(-\beta e_{ik}(W))}{\sum_{l=1}^{K} \exp(-\beta e_{il}(W))}
$$

  2. keeping the $\mu_{ik}$ constant, solve for $W$ (maximization phase)

$$
\sum_{i=1}^{N} \sum_{k=1}^{K} \mu_{ik} \nabla_W e_{ik}(W) = 0
$$

  3. loop to 1 until convergence
Fixed point scheme

- a simple strategy to solve $\nabla F_\beta(W) = 0$
- starting from a random value of $W$:
  1. compute (expectation phase)

$$
\mu_{ik} = \frac{\exp(-\beta e_{ik}(W))}{\sum_{l=1}^{K} \exp(-\beta e_{il}(W))}
$$

  2. keeping the $\mu_{ik}$ constant, solve for $W$ (maximization phase)

$$
\sum_{i=1}^{N} \sum_{k=1}^{K} \mu_{ik} \nabla_w e_{ik}(W) = 0
$$

  3. loop to 1 until convergence
- generally converges to a minimum of $F_\beta(W)$ (for a fixed $\beta$)
Vector quantization

- if \( e_{ik}(W) = \|x_i - w_k\|^2 \) (vector quantization),

\[
\nabla_{w_k} F_\beta(W) = 2 \sum_{i=1}^{N} \mu_{ik}(W)(w_k - x_i),
\]

with

\[
\mu_{ik}(W) = \frac{\exp(-\beta \|x_i - w_k\|^2)}{\sum_{l=1}^{K} \exp(-\beta \|x_i - w_l\|^2)}
\]

- setting \( \nabla F_\beta(W) = 0 \) leads to

\[
w_k = \frac{1}{\sum_{i=1}^{N} \mu_{ik}(W) \sum_{i=1}^{N} \mu_{ik}(W)x_i}
\]
Links with the K-means

- if $\mu_{il} = \delta_{k(i)=l}$, we obtained the prototype update rule of the K-means

- in the K-means, we have

$$k(i) = \arg \min_{1 \leq l \leq K} \|x_i - w_k\|^2$$

- in deterministic annealing, we have

$$\mu_{ik}(W) = \frac{\exp(-\beta\|x_i - w_k\|^2)}{\sum_{l=1}^{K} \exp(-\beta\|x_i - w_l\|^2)}$$

- this is a soft minimum version of the crisp rule of the k-means
Links with EM

- isotropic Gaussian mixture with a unique and fixed variance $\epsilon$

$$p_k(x|w_k) = \frac{1}{(2\pi \epsilon)^{p/2}} e^{-\frac{1}{2\epsilon} \|x_i - w_k\|^2}$$

- given the mixing coefficients $\delta_k$, responsibilities are

$$P(x_i \in C_k|x_i, W) = \frac{\delta_k e^{-\frac{1}{2\epsilon} \|x_i - w_k\|^2}}{\sum_{l=1}^{K} \delta_l e^{-\frac{1}{2\epsilon} \|x_i - w_l\|^2}}$$

- identical update rules for $W$

- $\beta$ can be seen has an inverse variance: quantify the uncertainty about the clustering results
So far...

- if \( E(M, W) = \sum_{i=1}^{N} \sum_{k=1}^{K} M_{ik} e_{ik}(W) \), one tries to reach
  \[ \min_{M \in \mathcal{M}, W \in \mathbb{R}^p} E(M, W) \]
  using
  \[
  F_\beta(W) = -\frac{1}{\beta} \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \exp(-\beta e_{ik}(W))
  \]
  - \( F_\beta(W) \) is smooth
  - a fixed point EM-like scheme can be used to minimize \( F_\beta(W) \) for a fixed \( \beta \)
So far...

- if \( E(M, W) = \sum_{i=1}^{N} \sum_{k=1}^{K} M_{ik} e_{ik}(W) \), one tries to reach
  \( \min_{M \in \mathcal{M}, W \in \mathbb{R}^p} E(M, W) \) using
  \[
  F_\beta(W) = -\frac{1}{\beta} \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \exp(-\beta e_{ik}(W))
  \]

- \( F_\beta(W) \) is smooth
- a fixed point EM-like scheme can be used to minimize \( F_\beta(W) \) for a fixed \( \beta \)
- but:
  - the k-means algorithm is also a EM like algorithm: it does not reach a global optimum
  - how handle to handle \( \beta \rightarrow \infty \)?
Outline

Introduction

Mixed problems
  Soft minimum
  Computing the soft minimum
  Evolution of $\beta$

Deterministic Annealing
  Annealing
  Maximum entropy
  Phase transitions
  Mass constrained deterministic annealing

Combinatorial problems
  Expectation approximations
  Mean field annealing
  In practice
Limit case $\beta \to 0$

- limit behavior of the EM scheme:
  - $\mu_{ik} = \frac{1}{K}$ for all $i$ and $k$ (and $W$!)
  - $W$ is therefore a solution of
    \[
    \sum_{i=1}^{N} \sum_{k=1}^{K} \nabla_W e_{ik}(W) = 0
    \]
  - no iteration needed
Limit case $\beta \to 0$

- limit behavior of the EM scheme:
  - $\mu_{ik} = \frac{1}{K}$ for all $i$ and $k$ (and $W$!)
  - $W$ is therefore a solution of
    \[
    \sum_{i=1}^{N} \sum_{k=1}^{K} \nabla_W e_{ik}(W) = 0
    \]
  - no iteration needed

- vector quantization:
  - we have
    \[
    w_k = \frac{1}{N} \sum_{i=1}^{N} x_i
    \]
  - each prototype is the center of mass of the data: only one cluster!

- in general the case $\beta \to 0$ is easy (unique minimum under mild hypotheses)
Example

Clustering in 2 clusters elements from $\mathbb{R}$

back to the example
Example

original problem
Example

simple soft minimum with $\beta = 10^{-1}$
Example

more complex soft minimum with $\beta = 0.5$
Increasing $\beta$

- when $\beta$ increases, $F_\beta(W)$ converges to $\min_{M \in \mathcal{M}} E(M, W)$
- optimizing $F_\beta(W)$ becomes more and more complex:
  - local minima
  - rougher and rougher: $F_\beta(W)$ remains $C^1$ but with large values for the gradient at some points

- path following strategy (homotopy):
  - for an increasing series $\beta_l$ with $\beta_0 = 0$
  - initialize $W^0 = \arg\min_W F_0(W)$ for $\beta = 0$ by solving
    $$\sum_{i=1}^N \sum_{k=1}^K \nabla W e_{ik}(W) = 0$$
  - compute $W^l$ via the EM scheme for $\beta_l$ starting from $W^{l-1}$
- a similar strategy is used in interior point algorithms
Increasing $\beta$

- when $\beta$ increases, $F_\beta(W)$ converges to $\min_{M \in \mathcal{M}} E(M, W)$
- optimizing $F_\beta(W)$ becomes more and more complex:
  - local minima
  - rougher and rougher: $F_\beta(W)$ remains $C^1$ but with large values for the gradient at some points
- path following strategy (homotopy):
  - for an increasing series $(\beta_l)_l$ with $\beta_0 = 0$
  - initialize $W_0^* = \arg \min_w F_0(W)$ for $\beta = 0$ by solving
    $$\sum_{i=1}^N \sum_{k=1}^K \nabla_w e_{ik}(W) = 0$$
  - compute $W_l^*$ via the EM scheme for $\beta_l$ starting from $W_{l-1}^*$
Increasing $\beta$

- when $\beta$ increases, $F_\beta(W)$ converges to $\min_{M \in \mathcal{M}} E(M, W)$
- optimizing $F_\beta(W)$ becomes more and more complex:
  - local minima
  - rougher and rougher: $F_\beta(W)$ remains $C^1$ but with large values for the gradient at some points
- path following strategy (homotopy):
  - for an increasing series $(\beta_l)_l$ with $\beta_0 = 0$
  - initialize $W_0^* = \arg \min_W F_0(W)$ for $\beta = 0$ by solving
    \[ \sum_{i=1}^N \sum_{k=1}^K \nabla_W e_{ik}(W) = 0 \]
  - compute $W_l^*$ via the EM scheme for $\beta_l$ starting from $W_{l-1}^*$
- a similar strategy is used in interior point algorithms
Example

Clustering in 2 classes of 6 elements from $\mathbb{R}$
Behavior of $F_\beta(W)$

$$\min_{M \in \mathcal{M}} E(M, W)$$
Behavior of $F_\beta(W)$

$F_\beta(W), \beta = 10^{-2}$
Behavior of $F_\beta(W)$

$F_\beta(W), \beta = 10^{-1.75}$
Behavior of $F_{\beta}(W)$

$F_{\beta}(W), \beta = 10^{-1.5}$
Behavior of $F_\beta(W)$

$F_\beta(W)$, $\beta = 10^{-1.25}$
Behavior of $F_{\beta}(W)$

$F_{\beta}(W), \beta = 10^{-1}$
Behavior of $F_\beta(W)$

$f_\beta(W), \beta = 10^{-0.75}$
Behavior of $F_\beta(W)$

$F_\beta(W)$, $\beta = 10^{-0.5}$
Behavior of $F_\beta(W)$

$F_\beta(W), \beta = 10^{-0.25}$
Behavior of $F_\beta(W)$

$F_\beta(W), \beta = 1$
Behavior of $F_\beta(W)$

$$\min_{M \in \mathcal{M}} E(M, W)$$
One step closer to the final algorithm

- given an increasing series \((\beta_l)_l\) with \(\beta_0 = 0\):
  1. compute \(W_0^*\) such that \(\sum_{i=1}^{N} \sum_{k=1}^{K} \nabla_w e_{ik}(W_0^*) = 0\)
  2. for \(l = 1\) to \(L\):
     2.1 initialize \(W_l^*\) to \(W_{l-1}^*\)
     2.2 compute \(\mu_{ik} = \frac{\exp(-\beta e_{ik}(W_l^*))}{\sum_{l=1}^{K} \exp(-\beta e_{il}(W_l^*))}\)
     2.3 update \(W_l^*\) such that \(\sum_{i=1}^{N} \sum_{k=1}^{K} \mu_{ik} \nabla_w e_{ik}(W_l^*) = 0\)
     2.4 good back to 2.2 until convergence
  3. use \(W_L^*\) as an estimation of \(\arg\min_{W \in \mathbb{R}^p} (\min_{M \in \mathcal{M}} E(M, W))\)

- this is (up to some technical details) the deterministic annealing algorithm for minimizing \(E(M, W) = \sum_{i=1}^{N} \sum_{k=1}^{K} M_{ik} e_{ik}(W)\)
What’s next?

- general topics:
  - why the name *annealing*?
  - why should that work? (and other philosophical issues)

- specific topics:
  - technical “details” (e.g., annealing schedule)
  - generalization:
    - non additive criteria
    - general combinatorial problems
Outline

Introduction

Mixed problems
  Soft minimum
  Computing the soft minimum
  Evolution of $\beta$

Deterministic Annealing
  Annealing
  Maximum entropy
  Phase transitions
  Mass constrained deterministic annealing

Combinatorial problems
  Expectation approximations
  Mean field annealing
  In practice
Annealing

from Wikipedia:

Annealing is a heat treatment wherein a material is altered, causing changes in its properties such as strength and hardness. It is a process that produces conditions by heating to above the re-crystallization temperature and maintaining a suitable temperature, and then cooling.
Annealing

▶ from Wikipedia:

Annealing is a heat treatment wherein a material is altered, causing changes in its properties such as strength and hardness. It is a process that produces conditions by heating to above the re-crystallization temperature and maintaining a suitable temperature, and then cooling.

▶ in our context, we’ll see that

▶ \( T = \frac{1}{k_B \beta} \) acts as a temperature and models some thermal agitation
▶ \( E(W, M) \) is the energy of a configuration the system while \( F_\beta(W) \) corresponds to the free energy of the system at a given temperature
▶ increasing \( \beta \) reduces the thermal agitation
Simulated Annealing

- a classical combinatorial optimization algorithm for computing \( \min_{M \in \mathcal{M}} E(M) \)
- given an increasing series \((\beta_l)_l\)
  1. choose a random initial configuration \(M_0\)
  2. for \(l = 1\) to \(L\)
     2.1 take a small random step from \(M_{l-1}\) to build \(M^c_l\) (e.g., change the cluster of an object)
     2.2 if \(E(M^c_l) < E(M_{l-1})\) then set \(M_l = M^c_l\)
     2.3 else set \(M_l = M^c_l\) with probability \(\exp(-\beta_l(E(M^c_l) - E(M_{l-1})))\)
     2.4 else set \(M_l = M_{l-1}\) in the other case
  3. use \(M_L\) as an estimation of \(\arg\min_{M \in \mathcal{M}} E(M)\)
- naive vision:
  - always accept improvement
  - “thermal agitation” allows to escape local minima
Statistical physics

- consider a system with state space $\mathcal{M}$
- denote $E(M)$ the energy of the system when in state $M$
- at thermal equilibrium with the environment, the probability for the system to be in state $M$ is given by the Boltzmann (Gibbs) distribution

$$P_T(M) = \frac{1}{Z_T} \exp \left( - \frac{E(M)}{k_B T} \right)$$

with $T$ the temperature, $k_B$ Boltzmann’s constant, and

$$Z_T = \sum_{M \in \mathcal{M}} \exp \left( - \frac{E(M)}{k_B T} \right)$$
Back to annealing

- **physical analogy:**
  - maintain the system at thermal equilibrium:
    - set a temperature
    - wait for the system to settle at this temperature
  - slowly decrease the temperature:
    - works well in real systems (e.g., crystallization)
    - allows the system to explore the state space

- **computer implementation:**
  - direct computation of $P_T(M)$ (or related quantities)
  - sampling from $P_T(M)$
Simulated Annealing
Revisited

- simulated annealing samples from $P_T(M)$!
- more precisely: the asymptotic distribution of $M_i$ for a fixed $\beta$ is given by $P_1/k_b\beta$
Simulated Annealing
Revisited

- simulated annealing samples from $P_T(M)$!
- more precisely: the asymptotic distribution of $M_i$ for a fixed $\beta$ is given by $P_1/k_b\beta$
- how?
  - Metropolis-Hastings Markov Chain Monte Carlo
  - principle:
    - $P$ is the target distribution
    - $Q(.|.)$ is the proposal distribution (sampling friendly)
    - start with $x_t$, get $x'$ from $Q(x|x_t)$
    - set $x^{t+1}$ to $x'$ with probability
      $$\min\left(1, \frac{P(x')Q(x_t|x')}{P(x_t)Q(x'|x_t)}\right)$$
    - keep $x^{t+1} = x^t$ when this fails
Simulated Annealing
Revisited

- in SA, $Q$ is the random local perturbation
- major feature of Metropolis-Hastings MCMC:

$$\frac{P_T(M')}{P_T(M)} = \exp\left(-\frac{E(M') - E(M)}{k_B T}\right)$$

$Z_T$ is not needed
- underlying assumption, symmetric proposals

$$Q(x^t|x') = Q(x'|x^t)$$

- rationale:
  - sampling directly from $P_T(M)$ for small $T$ is difficult
  - track likely area during cooling
Mixed problems

- Simulated Annealing applies to combinatorial problems

- in mixed problems, this would means removing the continuous part

\[
\min_{M \in \mathcal{M}, W \in \mathbb{R}^p} E(M, W) = \min_{M \in \mathcal{M}} \left( M \mapsto \min_{W \in \mathbb{R}^p} E(M, W) \right)
\]

- in the vector quantization example

\[
E(M) = \sum_{i=1}^{N} \sum_{k=1}^{K} M_{ik} \left\| x_i - \frac{1}{\sum_{j=1}^{N} M_{jk}} \sum_{j=1}^{N} M_{jk} x_j \right\|^2
\]

- no obvious direct relation to the proposed algorithm
thermodynamic (free) energy: the useful energy available in a system (a.k.a., the one that can be extracted to produce work)

Helmholtz (free) energy is $F_T = U - TS$, where $U$ is the internal energy of the system and $S$ its entropy

one can shown that

$$F_T = -k_B T \ln Z_T$$

the natural evolution of a system is to reduce its free energy

deterministic annealing mimicks the cooling process of a system by tracking the evolution of the minimal free energy
Gibbs distribution

- still no use of $P_T$...
Gibbs distribution

- still no use of $P_T$...
- important property:
  - $f$ a function defined on $\mathcal{M}$
    
    $$\mathbb{E}_{P_T}(f(M)) = \frac{1}{Z_T} \sum_{M \in \mathcal{M}} f(M) \exp \left( - \frac{E(M)}{k_B T} \right)$$

- then
  
  $$\lim_{T \to 0} \mathbb{E}_{P_T}(f(M)) = f(M^*),$$

  where $M^* = \arg\min_{M \in \mathcal{M}} E(M)$

- useful to track global information
Membership functions

- in the EM like phase, we have

\[ \mu_{ik} = \frac{\exp(-\beta e_{ik}(W))}{\sum_{l=1}^{K} \exp(-\beta e_{il}(W))} \]
Membership functions

- In the EM like phase, we have

\[ \mu_{ik} = \frac{\exp(-\beta e_{ik}(W))}{\sum_{l=1}^{K} \exp(-\beta e_{il}(W))} \]

- This does not come out of thin air:

\[ \mu_{ik} = \mathbb{E}_{P_{\beta,W}}(M_{ik}) = \sum_{M \in \mathcal{M}} M_{ik} P_{\beta,W}(M) \]

- \( \mu_{ik} \) is therefore the probability for \( x_i \) to belong to cluster \( k \) under the Gibbs distribution.
Membership functions

- in the EM like phase, we have

\[ \mu_{ik} = \frac{\exp(-\beta e_{ik}(W))}{\sum_{l=1}^{K} \exp(-\beta e_{il}(W))} \]

- this does not come out of thin air:

\[ \mu_{ik} = \mathbb{E}_{P_{\beta,W}}(M_{ik}) = \sum_{M \in \mathcal{M}} M_{ik} P_{\beta,W}(M) \]

- \( \mu_{ik} \) is therefore the probability for \( x_i \) to belong to cluster \( k \) under the Gibbs distribution

- at the limit \( \beta \to \infty \), the \( \mu_{ik} \) peak to Dirac like membership functions: they give the “optimal” partition
Example

Clustering in 2 classes of 6 elements from $\mathbb{R}$

\[
\begin{array}{cccccc}
1.00 & 2.00 & 3.00 & 7.00 & 8.25 \\
\end{array}
\]
Membership functions

[Graph showing membership functions with axes labeled $w_1$ and $w_2$, and a probability axis with values 0.0 to 1.0. The graph includes a point at the origin and bars at values 1, 2, 3, 7, 7.5, and 8.25 on the $x$-axis.]
Membership functions
Membership functions

![Contour plot and probability bars]

- **$w_1$** and **$w_2$** axes with values 2, 4, 6, 8.
- **Probability** axis with values 0.0, 0.2, 0.4, 0.6, 0.8, 1.0.
- Contour plot with concentric circles.
- Probability bars for $x$ values: 1, 2, 3, 7, 7.5, 8.25.
Membership functions

before phase transition
Membership functions

after phase transition
Membership functions
Membership functions

![Graph showing membership functions with coordinates w₁ and w₂ on the x-axis and y-axis respectively, and probability on the y-axis. The graph includes contour lines and probability bars for different values of x.]
Membership functions

![Diagram showing membership functions with axes labeled w1 and w2. The right side shows probability with values at x = 1, 2, 3, 7, 7.5, and 8.25. The plot contains concentric circles and a point marked with a red dot.]
Membership functions
Summary

- two principles:
  - physical systems tend to reach a state of minimal free energy at a given temperature
  - when the temperature is slowly decreased, a system tends to reach a well organized state

- annealing algorithms tend to reproduce this slow cooling behavior

- simulated annealing:
  - uses Metropolis Hasting MCMC to sample the Gibbs distribution
  - easy to implement but needs a very slow cooling

- deterministic annealing:
  - direct minimization of the free energy
  - computes expectations of global quantities with respect to the Gibbs distribution
  - aggressive cooling is possible, but $Z_T$ computation must be tractable
Outline

Introduction

Mixed problems
  Soft minimum
  Computing the soft minimum
  Evolution of $\beta$

Deterministic Annealing
  Annealing
  Maximum entropy
  Phase transitions
  Mass constrained deterministic annealing

Combinatorial problems
  Expectation approximations
  Mean field annealing
  In practice
Maximum entropy

- another interpretation/justification of DA
- reformulation of the problem: find a probability distribution on $\mathcal{M}$ which is “regular” and gives a low average error
Maximum entropy

- another interpretation/justification of DA
- reformulation of the problem: find a probability distribution on $\mathcal{M}$ which is “regular” and gives a low average error
- in other words: find $P_W$ such that
  - $\sum_{M \in \mathcal{M}} P_W(M) = 1$
  - $\sum_{M \in \mathcal{M}} E(W, M) P_W(M)$ is small
  - the entropy of $P_W$, $-\sum_{M \in \mathcal{M}} \ln P_W(M) P_W(M)$ is high
- entropy plays a regularization role
Maximum entropy

- another interpretation/justification of DA
- reformulation of the problem: find a probability distribution on $\mathcal{M}$ which is “regular” and gives a low average error
- in other words: find $P_W$ such that
  - $\sum_{M \in \mathcal{M}} P_W(M) = 1$
  - $\sum_{M \in \mathcal{M}} E(W, M) P_W(M)$ is small
  - the entropy of $P_W$, $-\sum_{M \in \mathcal{M}} \ln P_W(M) P_W(M)$ is high
- entropy plays a regularization role
- this leads to the minimization of

$$\sum_{M \in \mathcal{M}} E(W, M) P_W(M) + \frac{1}{\beta} \sum_{M \in \mathcal{M}} \ln P_W(M) P_W(M)$$

where $\beta$ sets the trade-off between fitting and regularity
Maximum entropy

- some calculations show that the minimum over $P_W$ is given by

$$P_{\beta,W}(M) = \frac{1}{Z_{\beta,W}} \exp(-\beta E(M, W)),$$

with

$$Z_{\beta,W} = \sum_{M \in \mathcal{M}} \exp(-\beta E(M, W)).$$

- plugged back into the optimization criterion, we end up with the soft minimum

$$F_{\beta}(W) = -\frac{1}{\beta} \ln \sum_{M \in \mathcal{M}} \exp(-\beta E(M, W)).$$
So far...

- three derivations of the deterministic annealing:
  - soft minimum
  - thermodynamic inspired annealing
  - maximum entropy principle
- all of them boil down to two principles:
  - replace the crisp optimization problem in the $\mathcal{M}$ space by a soft one
  - track the evolution of the solution when the crispness of the approximation increases
- now the technical details...
Outline

Introduction

Mixed problems
  Soft minimum
  Computing the soft minimum
  Evolution of $\beta$

Deterministic Annealing
  Annealing
  Maximum entropy
  Phase transitions
  Mass constrained deterministic annealing

Combinatorial problems
  Expectation approximations
  Mean field annealing
  In practice
Fixed point

- remember the EM phase:

\[ \mu_{ik}(W) = \frac{\exp(-\beta \| x_i - w_k \|^2)}{\sum_{l=1}^{K} \exp(-\beta \| x_i - w_l \|^2)} \]

\[ w_k = \frac{1}{\sum_{i=1}^{N} \mu_{ik}(W)} \sum_{i=1}^{N} \mu_{ik}(W)x_i \]

- this is a fixed point method, i.e., \( W = U_\beta(W) \) for a data dependent \( U_\beta \)

- the fixed point is generally stable, except during a phase transition
Phase transition

before phase transition
Phase transition

after phase transition
Stability
Stability

- unstable fixed point:
  - even during a phase transition, the exact previous fixed point can remain a fixed point
  - but we want the transition to take place: this is a new cluster birth!
  - fortunately, the previous fixed point is unstable during a phase transition
Stability

- Unstable fixed point:
  - Even during a phase transition, the exact previous fixed point can remain a fixed point.
  - But we want the transition to take place: this is a new cluster birth!
  - Fortunately, the previous fixed point is unstable during a phase transition.

- Modified EM phase:
  1. Initialize $W_{i}^{*}$ to $W_{i-1}^{*} + \epsilon$
  2. Compute $\mu_{ik} = \frac{\exp(-\beta e_{ik}(W_{i}^{*}))}{\sum_{i=1}^{K} \exp(-\beta e_{il}(W_{i}^{*}))}$
  3. Update $W_{i}^{*}$ such that $\sum_{i=1}^{N} \sum_{k=1}^{K} \mu_{ik} \nabla_{W} e_{ik}(W_{i}^{*}) = 0$
  4. Good back to 2.2 until convergence.

- The noise is important to prevent missing phase transition.
Annealing strategy

- neglecting symmetries, there are $K - 1$ phase transitions for $K$ clusters
- tracking $W^*$ between transitions is easy
Annealing strategy

- neglecting symmetries, there are $K - 1$ phase transitions for $K$ clusters
- tracking $W^*$ between transitions is easy
- trade-off:
  - slow annealing: long running time but no transition is missed
  - fast annealing: quicker but with higher risk to miss a transition
Annealing strategy

- neglecting symmetries, there are $K - 1$ phase transitions for $K$ clusters
- tracking $W^*$ between transitions is easy
- trade-off:
  - slow annealing: long running time but no transition is missed
  - fast annealing: quicker but with higher risk to miss a transition
- but critical temperatures can be computed:
  - via a stability analysis of fixed points
  - corresponds to a linear approximation of the fixed point equation $U_\beta$ around a fixed point
Critical temperatures

- for vector quantization
- first temperature:
  - the fixed point stability related to the variance of the data
  - the critical $\beta$ is $1/2\lambda_{\text{max}}$ where $\lambda_{\text{max}}$ is the largest eigenvalue of the covariance matrix of the data

...
Critical temperatures

▶ for vector quantization
▶ first temperature:
  ▶ the fixed point stability related to the variance of the data
  ▶ the critical $\beta$ is $1/2\lambda_{\text{max}}$ where $\lambda_{\text{max}}$ is the largest eigenvalue of the covariance matrix of the data
▶ in general:
  ▶ the stability of the whole system is related to the stability of each cluster
  ▶ the $\mu_{ik}$ play the role of membership functions for each cluster
  ▶ the critical $\beta$ for cluster $k$ is $1/2\lambda_k^{\text{max}}$ where $\lambda_k^{\text{max}}$ is the largest eigenvalue of

$$
\sum_{i} \mu_{ik}(x_i - w_k)(x_i - w_k)^T
$$
A possible final algorithm

1. initialize $W^1 = \frac{1}{N} \sum_{i=1}^{N} x_i$

2. for $l = 2$ to $K$:
   2.1 compute the critical $\beta_l$
   2.2 initialize $W^l$ to $W^{l-1}$
   2.3 for $t$ values of $\beta$ around $\beta_l$
      2.3.1 add some small noise to $W^l$
      2.3.2 compute $\mu_{ik} = \exp(-\beta \|x_i - W^l_k\|^2) \sum_{t=1}^{K} \exp(-\beta \|x_i - W^l_t\|^2)$
      2.3.3 compute $W^l_k = \frac{1}{\sum_{i=1}^{N} \mu_{ik}} \sum_{i=1}^{N} \mu_{ik} x_i$
      2.3.4 good back to 2.3.2 until convergence

3. use $W^K$ as an estimation of
   $\arg\min_{W \in \mathbb{R}^p} (\min_{M \in \mathcal{M}} E(M, W))$ and $\mu_{ik}$ as the corresponding $M$
Computational cost

- $N$ observations in $\mathbb{R}^p$ and $K$ clusters
- prototype storage: $Kp$
- $\mu$ matrix storage: $NK$
- one EM iteration costs: $O(NKp)$
- we need at most $K - 1$ full EM runs: $O(NK^2p)$
- compared to the K-means:
  - more storage
  - no simple distance calculation tricks: all distances must be computed
  - roughly corresponds to $K - 1$ k-means, not taking into account the number of distinct $\beta$ considered during each phase transition
- neglecting the eigenvalue analysis (in $O(p^3))$...
Outline

Introduction

Mixed problems
  Soft minimum
  Computing the soft minimum
  Evolution of $\beta$

Deterministic Annealing
  Annealing
  Maximum entropy
  Phase transitions
  Mass constrained deterministic annealing

Combinatorial problems
  Expectation approximations
  Mean field annealing
  In practice
Collapsed prototypes

- when $\beta \to 0$, only one cluster
- waste of computational resources: $K$ identical calculations
- this is a general problem:
  - each phase transition adds new clusters
  - prior to that prototypes are collapsed
  - but we need them to track cluster birth...

$P(x_i \in C_k | x_i, \omega) = \delta_{k \epsilon} - \frac{1}{2} \epsilon \|x_i - w_k\|^2$

prototypes are weighted by the mixing coefficients $\delta$.
Collapsed prototypes

- when $\beta \to 0$, only one cluster
- waste of computational resources: $K$ identical calculations
- this is a **general problem**:
  - each phase transition adds new clusters
  - prior to that prototypes are collapsed
  - but we need them to track cluster birth...
- a related problem: uniform cluster weights
- in a Gaussian mixture

\[
P(x_i \in C_k | x_i, W) = \frac{\delta_k e^{-\frac{1}{2\epsilon} \|x_i-w_k\|^2}}{\sum_{l=1}^{K} \delta_l e^{-\frac{1}{2\epsilon} \|x_i-w_l\|^2}}
\]

prototypes are **weighted** by the mixing coefficients $\delta$
Mass constrained DA

- we introduce prototype weights $\delta_k$ and plug them into the soft minimum formulation

$$F_\beta(W, \delta) = -\frac{1}{\beta} \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \delta_k \exp(-\beta e_{ik}(W))$$

- $F_\beta$ is minimized under the constraint $\sum_{k=1}^{K} \delta_k = 1$
- this leads to

$$\mu_{ik}(W) = \frac{\delta_k \exp(-\beta \| x_i - w_k \|^2)}{\sum_{l=1}^{K} \delta_l \exp(-\beta \| x_i - w_l \|^2)}$$

$$\delta_k = \sum_{i=1}^{N} \mu_{ik}(W) / N$$

$$w_k = N \sum_{i=1}^{N} \mu_{ik}(W) x_i / \delta_k$$
MCDA

- increases the similarity with EM for Gaussian mixtures:
  - exactly the same algorithm for a fixed $\beta$
  - isotropic Gaussian mixture with variance $\frac{1}{2\beta}$

- however:
  - the variance is generally a parameter in the Gaussian mixtures
  - different goals: minimal distortion versus maximum likelihood
  - no support for other distributions in DA
Cluster birth monitoring

- avoid wasting computational resources:
- consider a cluster $C_k$ with a prototype $w_k$
- prior a phase transition in $C_k$:
  - duplicate $w_k$ into $w'_k$
  - apply some noise to both prototypes
  - split $\delta_k$ into $\frac{\delta_k}{2}$ and $\frac{\delta'_k}{2}$
  - apply the EM like algorithm and monitor $\|w_k - w'_k\|$
  - accept the new cluster if $\|w_k - w'_k\|$ becomes “large”
- two strategies:
  - eigenvalue based analysis (costly, but quite accurate)
  - opportunistic: always maintain duplicate prototypes and promote diverging ones
Final algorithm

1. initialize \( W^1 = \frac{1}{N} \sum_{i=1}^{N} x_i \)

2. for \( l = 2 \) to \( K \):
   2.1 compute the critical \( \beta_l \)
   2.2 initialize \( W^l \) to \( W^{l-1} \)
   2.3 duplicate the prototype of the critical cluster and split the associated weight
   2.4 for \( t \) values of \( \beta \) around \( \beta_l \)
      2.4.1 add some small noise to \( W^l \)
      2.4.2 compute \( \mu_{ik} = \frac{\delta_k \exp(-\beta \| x_i - W^l_k \|^2)}{\sum_{t=1}^{K} \delta_t \exp(-\beta \| x_i - W^l_t \|^2)} \)
      2.4.3 compute \( \delta_k = \sum_{i=1}^{N} \mu_{ik}(W)/N \)
      2.4.4 compute \( W^l_k = \frac{N}{\delta_k} \sum_{i=1}^{N} \mu_{ik} x_i \)
      2.4.5 good back to 2.4.2 until convergence

3. use \( W^K \) as an estimation of \( \arg \min_{W \in \mathbb{R}^p} (\min_{M \in \mathcal{M}} E(M, W)) \) and \( \mu_{ik} \) as the corresponding \( M \)
A simple dataset
Error evolution and cluster births
Clusters
Clusters

![Cluster visualization with scattered data points on a 2D graph with axes X1 and X2. The data points are clustered into several distinct groups.]}
Clusters

-2 0 2 4 6
-2 −1 0 1 2 3 4
X1
X2
Clusters

![Cluster visualization](image-url)
Clusters
Clusters
Clusters

![Data Clusters Plot](image)

The plot illustrates clusters in a two-dimensional space defined by variables $X_1$ and $X_2$. The data points are scattered across the plot, forming distinct clusters. The red dots represent the centroids of the clusters.
Clusters
Clusters
Clusters

\[ X_1 \]
\[ X_2 \]
Clusters
Summary

- Deterministic annealing addresses combinatorial optimization by smoothing the cost function and tracking the evolution of the solution while the smoothing progressively vanishes.

- Motivated by:
  - Heuristic (soft minimum)
  - Statistical physics (Gibbs distribution)
  - Information theory (maximal entropy)

- In practice:
  - Rather simple multiple EM like algorithm
  - A bit tricky to implement (phase transition, noise injection, etc.)
  - Excellent results (frequently better than those obtained by K-means)
Outline

Introduction

Mixed problems
   Soft minimum
   Computing the soft minimum
   Evolution of $\beta$

Deterministic Annealing
   Annealing
   Maximum entropy
   Phase transitions
   Mass constrained deterministic annealing

Combinatorial problems
   Expectation approximations
   Mean field annealing
   In practice
Combinatorial only

- the situation is quite different for a combinatorial problem
  \[
  \arg \min_{M \in \mathcal{M}} E(M)
  \]

- no soft minimum heuristic

- no direct use of the Gibbs distribution \( P_T \); one need to rely on \( \mathbb{E}_{P_T} (f(M)) \) for interesting \( f \)

- calculation of \( P_T \) is not tractable:
  - tractability is related to independence
  - if
    \[
    E(M) = \sum_{M_1} \ldots \sum_{M_D} (E(M_1) + \ldots + E(M_D))
    \]
    then independent optimization can be done on each variable...
Graph clustering

- a non oriented graph with $N$ nodes and weight matrix $A$ ($A_{ij}$ is the weight of the connection between nodes $i$ and $j$)
- maximal **modularity** clustering:
  - node degree $k_i = \sum_{j=1}^{N} A_{ij}$, total weight $m = \frac{1}{2} \sum_{i,j} A_{ij}$
  - null model $P_{ij} = \frac{k_i k_j}{2m}$, $B_{ij} = \frac{1}{2m} (A_{ij} - P_{ij})$
  - assignment matrix $M$
  - Modularity of the clustering
    \[
    \text{Mod}(M) = \sum_{i,j} \sum_{k} M_{ik} M_{jk} B_{ij}
    \]
- difficulty: coupling $M_{ik} M_{jk}$ (non-linearity)
- in the sequel $E(M) = -\text{Mod}(M)$
choose some interesting statistics:
- $\mathbb{E}_{P_T}(f(M))$
- for instance $\mathbb{E}_{P_T}(M_{ik})$ for clustering

compute an approximation of $\mathbb{E}_{P_T}(f(M))$ at high temperature $T$

track the evolution of the approximation while lowering $T$

use $\lim_{T \to 0} \mathbb{E}_{P_T}(f(M)) = f(\arg \min_{M \in \mathcal{M}} E(M))$ to claim that the approximation converges to the interesting limit

rationale: approximating $\mathbb{E}_{P_T}(f(M))$ for a low $T$ is probably difficult, we use the path following strategy to ease the task
Example
Example

\[ \beta = 10^{-1} \]
Example

$\beta = 10^{-0.75}$
Example

\[ \beta = 10^{-0.5} \]

\[ \text{probability} \]

\[ x \]

\[ \text{probability} \]

\[ 0.000 \ 0.005 \ 0.010 \ 0.015 \ 0.020 \ 0.025 \]

\[ -1.0 \ -0.5 \ 0.0 \ 0.5 \ 1.0 \]
Example

\[ \beta = 10^{-0.25} \]
Example

\[ \beta = 10^0 \]
Example

\[ \beta = 10^{0.25} \]
Example

$\beta = 10^{0.5}$
\[ \beta = 10^{0.75} \]
Example

\[ \beta = 10^1 \]

\[ x \]

\[ \text{probability} \]

\[ -1.0 \quad -0.5 \quad 0.0 \quad 0.5 \quad 1.0 \]

\[ 0.0 \quad 0.1 \quad 0.2 \quad 0.3 \quad 0.4 \quad 0.5 \]
Example

$\beta = 10^{1.25}$

- $x$
- probability

$\beta = 10^{1.25}$
Example

$\beta = 10^{1.5}$
Example

\[ \beta = 10^{1.75} \]
Example

\[ \beta = 10^2 \]

The diagram shows a probability distribution with a range from -1.0 to 1.0 on the x-axis and a probability range from 0.0 to 1.0 on the y-axis. The value of \( \beta \) is indicated at 0.5 on the x-axis, highlighting a specific point of interest in the distribution.
Example
Expectation approximation

- computing

\[ \mathbb{E}_{P_Z}(f(Z)) = \int f(Z) dP_Z \]

has many applications
- for instance, in machine learning:
  - performance evaluation (generalization error)
  - EM for probabilistic models
  - Bayesian approaches
- two major numerical approaches:
  - sampling
  - variational approximation
Sampling

- (strong) law of large numbers

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} f(z_i) = \int f(Z) dP_Z
\]

if the \( z_i \) are independent samples of \( Z \)

- many extensions, especially to dependent samples (slower convergence)

- MCMC methods: build a Markov chain with stationary distribution \( P_Z \) and take the average of \( f \) on a trajectory

- applied to \( \mathbb{E}_{P_T}(f(M)) \): this is simulated annealing!
Sampling

- (strong) law of large numbers

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} f(z_i) = \int f(Z) \, dP_Z
\]

if the \( z_i \) are independent samples of \( Z \)

- many extensions, especially to dependent samples (slower convergence)

- MCMC methods: build a Markov chain with stationary distribution \( P_Z \) and take the average of \( f \) on a trajectory

- applied to \( \mathbb{E}_{P_T}(f(M)) \): this is simulated annealing!
Variational approximation

- main idea:
  - the intractability of the calculation of $\mathbb{E}_{P_Z}(f(Z))$ is induced by the complexity of $P_Z$
  - let's replace $P_Z$ by a simpler distribution $Q_Z$...
  - ...and $\mathbb{E}_{P_Z}(f(Z))$ by $\mathbb{E}_{Q_Z}(f(Z))$

- calculus of variation:
  - optimization over functional spaces
  - in this context: choose an optimal $Q_Z$ in a space of probability measures $Q$ with respect to a goodness of fit criterion between $P_Z$ and $Q_Z$

- probabilistic context:
  - simple distributions: factorized distributions
  - quality criterion: Kullback-Leibler divergence
Variational approximation

- Assume $Z = (Z_1, \ldots, Z_D) \in \mathbb{R}^D$ and $f(Z) = \sum_{i=1}^{D} f_i(Z_i)$
- For a general $P_Z$, $f$ structure cannot be exploited in $\mathbb{E}_{P_Z}(f(Z))$
Variational approximation

- assume $Z = (Z_1, \ldots, Z_D) \in \mathbb{R}^D$ and $f(Z) = \sum_{i=1}^{D} f_i(Z_i)$
- for a general $P_Z$, $f$ structure cannot be exploited in $\mathbb{E}_{P_Z}(f(Z))$
- variational approximation
  - choose $Q$ a set of tractable distributions on $\mathbb{R}$
  - solve
  
  $$Q^* = \text{arg}\min_{Q=(Q_1 \times \ldots \times Q_D) \in \mathbb{Q}^D} \text{KL} (Q || P_Z)$$

  with
  
  $$\text{KL} (Q || P_Z) = - \int \ln \frac{dP_Z}{dQ} dQ$$

- approximate $\mathbb{E}_{P_Z}(f(Z))$ by

  $$\mathbb{E}_{Q^*}(f(Z)) = \sum_{i=1}^{D} \mathbb{E}_{Q_i^*}(f_i(Z_i))$$
Application to DA

- general form: \( P_\beta(M) = \frac{1}{Z_\beta} \exp(-\beta E(M)) \)
- intractable because \( E(M) \) is not linear in \( M = (M_1, \ldots, M_D) \)
Application to DA

- general form: \( P_\beta(M) = \frac{1}{Z_\beta} \exp(-\beta E(M)) \)
- intractable because \( E(M) \) is not linear in \( M = (M_1, \ldots, M_D) \)
- variational approximation:
  - replace \( E \) by \( F(W, M) \) defined by
    \[
    F(W, M) = \sum_{i=1}^{D} W_i M_i
    \]
  - use for \( Q_{W,\beta} \)
    \[
    Q_{W,\beta} = \frac{1}{Z_{W,\beta}} \exp(-\beta F(W, M))
    \]
  - optimize \( W \) via \( \text{KL}(Q_{W,\beta} \| P_\beta) \)
Fitting the approximation

- a complex expectation calculation is replaced by a new optimization problem
- we have

\[
\text{KL} \left( Q_{W,\beta} \parallel P_{\beta} \right) = \ln Z_{\beta} - \ln Z_{W,\beta} + \beta \mathbb{E}_{Q_{W,\beta}} \left( E(M) - F(W, M) \right)
\]

- tractable by factorization on $F$, except for $Z_{\beta}$
- solved by $\nabla_{W} \text{KL} \left( Q_{W,\beta} \parallel P_{\beta} \right) = 0$
  - tractable because $Z_{\beta}$ do not depend on $W$
  - generally solved via a fixed point approach (EM like again)
- similar to variational approximation in EM or Bayesian methods
Example

- in the case of (graph) clustering, we use

\[ F(W, M) = \sum_{i=1}^{N} \sum_{k=1}^{K} W_{ik} M_{ik} \]

- a crucial (general) property is that, when \( i \neq j \)

\[ \mathbb{E}_{Q_{W,\beta}} (M_{ik} M_{jl}) = \mathbb{E}_{Q_{W,\beta}} (M_{ik}) \mathbb{E}_{Q_{W,\beta}} (M_{jl}) \]

- a (long and boring) series of equations leads to the general mean field equations

\[ \frac{\partial \mathbb{E}_{Q_{W,\beta}} (E(M))}{\partial W_{jl}} = \sum_{k=1}^{K} \frac{\partial \mathbb{E}_{Q_{W,\beta}} (M_{jk})}{\partial W_{jl}} W_{jk}, \ \forall j, l. \]
Example

- classical EM like scheme to solve the mean field equations:
  1. we have
     \[ E_{Q,W,\beta}(M_{ik}) = \frac{\exp(-\beta W_{ik})}{\sum_{l=1}^{K} \exp(-\beta W_{il})} \]
  2. keep \( E_{Q,W,\beta}(M_{ik}) \) fixed and solve the simplified mean field equations for \( W \)
  3. update \( E_{Q,W,\beta}(M_{ik}) \) based on the new \( W \) and loop on 2 until convergence

- in the graph clustering case

\[ W_{jk} = 2 \sum_{i,j} E_{Q,W,\beta}(M_{ik}) B_{ij} \]
Summary

- deterministic annealing for combinatorial optimization
- given an objective function $E(M)$
  - choose a linear parametric approximation of $E(M)$, $F(W, M)$ with the associated distribution $Q_{W,\beta} = \frac{1}{Z_{W,\beta}} \exp(-\beta F(W, M))$
  - write the mean field equations, i.e.

\[
\nabla_W \left( -\ln Z_{W,\beta} + \beta \mathbb{E}_{Q_{W,\beta}} (E(M) - F(W, M)) \right) = 0
\]

- use a EM like algorithm to solve the equations:
  - given $W$, compute $\mathbb{E}_{Q_{W,\beta}} (M_{ik})$
  - given $\mathbb{E}_{Q_{W,\beta}} (M_{ik})$, solve the equations

- back to our classical questions:
  - why would that work?
  - how to do that in practice?
Outline

Introduction

Mixed problems
  Soft minimum
  Computing the soft minimum
  Evolution of $\beta$

Deterministic Annealing
  Annealing
  Maximum entropy
  Phase transitions
  Mass constrained deterministic annealing

Combinatorial problems
  Expectation approximations
  Mean field annealing
  In practice
Physical interpretation

- back to the (generalized) Helmholtz (free) energy

\[ F_\beta = U - \frac{S}{\beta} \]

- the free energy can be generalized to any distribution \( Q \) on the states, by

\[ F_{\beta,Q} = \mathbb{E}_Q(E(M)) - \frac{H(Q)}{\beta}, \]

where \( H(Q) = -\sum_{M \in M} Q(M) \ln Q(M) \) is the entropy of \( Q \)

- the Boltzmann-Gibbs distribution minimizes the free energy, i.e.

\[ F_T \leq F_{T,Q} \]
Physical interpretation

- In fact we have

\[ F_{T,Q} = F_T + \frac{1}{\beta} \text{KL}(Q||P), \]

where \( P \) is the Gibbs distribution

- Minimizing \( \text{KL}(Q||P) \geq 0 \) over \( Q \) corresponds to finding the best upper bound of the free energy on a class of distribution

- In other words: given the system states must be distributed according to a distribution in \( Q \), find the most stable distribution
Mean field

- $E(M)$ is difficult to handle because of coupling (dependencies) while $F(W, M)$ is linear and corresponds to a de-coupling in which dependencies are replaced by mean effects

  - example:
    - modularity

    $$E(M) = \sum_i \sum_k M_{ik} \left( \sum_j M_{jk} B_{ij} \right)$$

  - approximation

    $$F(W, M) = \sum_i \sum_k M_{ik} W_{ik}$$

- the complex influence of $M_{ij}$ on $E$ is replaced by a single parameter $W_{ik}$: the mean effect associated to a change in $M_{ij}$
Outline

Introduction

Mixed problems
  Soft minimum
  Computing the soft minimum
  Evolution of $\beta$

Deterministic Annealing
  Annealing
  Maximum entropy
  Phase transitions
  Mass constrained deterministic annealing

Combinatorial problems
  Expectation approximations
  Mean field annealing
  In practice
In practice

- homotopy again:
  - at high temperature \((\beta \to 0)\):
    - \[ E_{Q_{W,\beta}}(M_{ik}) = \frac{1}{K} \]
    - the fixed point equation is generally easy to solve; for instance in graph clustering
      \[ W_{jk} = \frac{2}{K} \sum_{i,j} B_{ij} \]
  - slowly decrease the temperature
  - use the mean field at the previous higher temperature as a starting point for the fixed point iterations
  - phase transitions:
    - stable versus unstable fixed points (eigenanalysis)
    - noise injection and mass constrained version
A graph clustering algorithm

1. initialize $W_{jk}^1 = \frac{2}{K} \sum_{i,j} B_{ij}$

2. for $l = 2$ to $K$:
   2.1 compute the critical $\beta_l$
   2.2 initialize $W^l$ to $W^{l-1}$
   2.3 for $t$ values of $\beta$ around $\beta_l$
      2.3.1 add some small noise to $W^l$
      2.3.2 compute $\mu_{ik} = \exp(-\beta W_{ik}) / \sum_{t=1}^{K} \exp(-\beta W_{it})$
      2.3.3 compute $W_{jk} = 2 \sum_{i,j} \mu_{ik} B_{ij}$
      2.3.4 good back to 2.3.2 until convergence

3. threshold $\mu_{ik}$ into an optimal partition of the original graph
   - mass constrained approach variant
   - stability based transition detection (slower annealing)
Karate

- clustering of a simple graph
- Zachary’s Karate club
Modularity

Evolution of the modularity during annealing
Phase transitions

- first phase: 2 clusters
Phase transitions

- **first phase**: 2 clusters
- **second phase**: 4 clusters
Karate

- clustering of a simple graph
- Zachary’s Karate club
- Four clusters
DA Roadmap

How to apply deterministic annealing to a combinatorial optimization problem $E(M)$?

1. define a linear mean field approximation $F(W, M)$

2. specialize the mean field equations, i.e. compute

$$\frac{\partial \mathbb{E}_{Q_{W, \beta}} (E(M))}{\partial W_{jl}} = \frac{\partial}{\partial W_{jl}} \left( \frac{1}{Z_{W, \beta}} \sum_M E(M) \exp(-\beta F(W, M)) \right)$$

3. identify a fixed point scheme to solve the mean field equations, using $\mathbb{E}_{Q_{W, \beta}} (M_i)$ as a constant if needed

4. wrap the corresponding EM scheme in an annealing loop: this is the basic algorithm

5. analyze the fixed point scheme to find stability conditions: this leads to an advanced annealing schedule
Summary

- Deterministic annealing addresses combinatorial optimization by solving a related but simpler problem and by tracking the evolution of the solution while the simpler problem converges to the original one.
- Motivated by:
  - Heuristic (soft minimum)
  - Statistical physics (Gibbs distribution)
  - Information theory (maximal entropy)
- Works mainly because of the solution following strategy (homotopy): do not solve a difficult problem from scratch, but rather starting from a good guess of the solution.
Summary

- **difficulties:**
  - boring and technical calculations (when facing a new problem)
  - a bit tricky to implement (phase transition, noise injection, etc.)
  - tends to be slow in practice: computing exp is very costly even on modern hardware (roughly 100 flop)

- **advantages:**
  - versatile
  - appears as a simple EM like algorithm embedded in an annealing loop
  - very interesting intermediate results
  - excellent final results in practice
