Machine Intelligence 2

3 Stochastic Optimization

Prof. Dr. Klaus Obermayer

Fachgebiet Neuronale Informationsverarbeitung (NI)

SS 2017

Stochastic Optimization

Simulated Annealing

Mean-Field Annealing

Stochastic optimization

Supervised & unsupervised learning $\rightarrow$ evaluation of cost function E^T

- real-valued arguments: gradient based techniques (e.g. ICA weights)
- discrete arguments: ?? (e.g. for cluster assignment)

$\Rightarrow$ simulated annealing

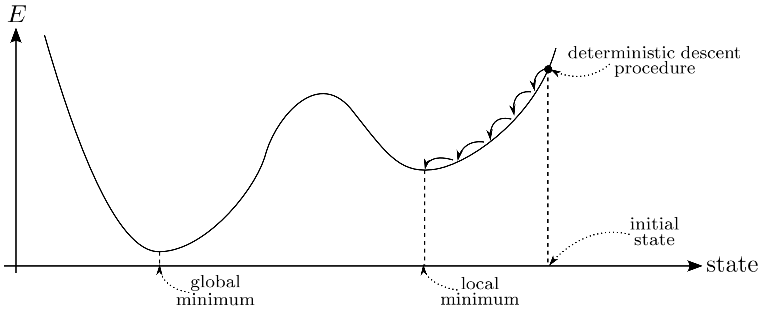
Setting

- discrete variables $s_i, \; i = 1, \ldots, N$ (e.g. $s_i \in \{+1, -1\}$ or $s_i \in \mathbb{N}$)
- short-hand notation: $\underline{\mathrm{s}}$ ("state") – often $\{\underline{\mathrm{s}}\}$ not a vector space (but called state space)
- **cost function:** $E : \underline{\mathbf{s}} \mapsto E_{(\underline{\mathbf{s}})} \in \mathbb{R}$ – not restricted to learning problems

Goal: find state $\underline{\mathrm{s}}^*$, such that:

$$E \overset{!}{=} \min \qquad \text{(desirable global minimum of } E\text{)}$$

Optimizing cost functions with local optima



- Deterministic descent may converge to local minima
- Grid-search, random search, multiple initializations
  $\rightsquigarrow$ *Simulated Annealing*

Simulated Annealing

History: "Naturalistic" stochastic optimization

$\rightsquigarrow$ mimicking freezing and crystallization

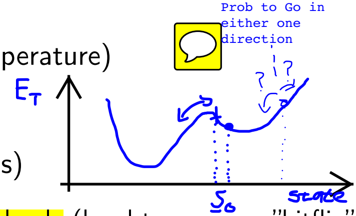
(atom configurations in crystals often close to global minima of the energy)

$\rightsquigarrow$ slow cooling (glass, unordered vs. crystal, ordered) $\Rightarrow$ annealing

$\Rightarrow$ slowly lower temperature while maintaining thermal equilibrium

$\Rightarrow$ computational temperature T or *noise parameter* $\beta = \frac{1}{T}$

Simulated Annealing



`initialization:` $\underline{\mathbf{s}}_0$, β_0 small ($\rightsquigarrow$ high temperature)

`BEGIN Annealing loop` $(t = 1, 2, \dots)$

  $\underline{\mathbf{s}}_t = \underline{\mathbf{s}}_{t-1}$ (initialization of inner loop)

  `BEGIN State update loop` (M iterations)

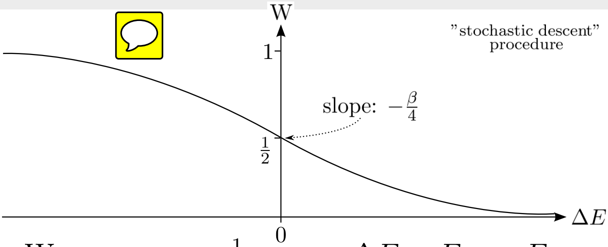
- `choose a new candidate state` $\underline{\mathbf{s}}$ randomly (local to $\underline{\mathbf{s}}_t$ – e.g. "bitflip")
- `calculate difference in cost:` $\Delta E = E_{(\underline{\mathbf{s}})} - E_{(\underline{\mathbf{s}}_t)}$
- `switch` $\underline{\mathbf{s}}_t$ `to` $\underline{\mathbf{s}}$ `with probability` $\mathrm{W}_{(\underline{\mathbf{s}}_t \to \underline{\mathbf{s}})} = \frac{1}{1+\exp(\beta_t \Delta E)}$ otherwise keep the previous state $\underline{\mathbf{s}}_t$

  `END State update loop`

  $\beta_t = \tau \beta_{t-1}$ $(\tau > 1 \implies$ increase of $\beta)$

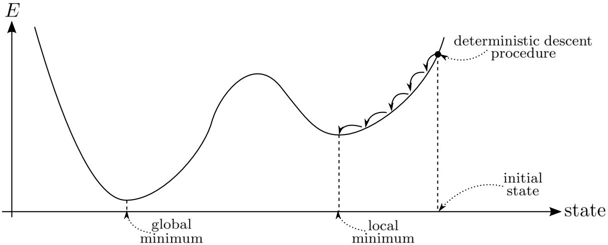
`END Annealing loop`

Transition probability



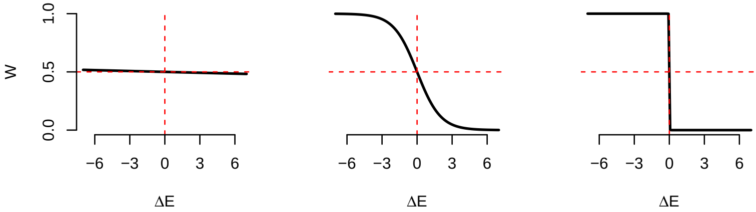
$$W_{(\underline{\mathbf{s}}_t \to \underline{\mathbf{s}})} = \frac{1}{1+\exp(\beta_t \Delta E)}, \quad \Delta E = E_{(\underline{\mathbf{s}})} - E_{(\underline{\mathbf{s}}_t)}$$

cost function with local optima:

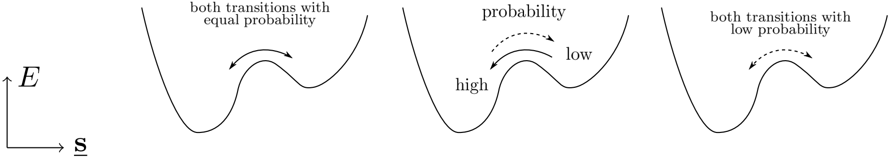


Annealing

limiting cases for high vs. low temperature:



low β (high temperature) intermediate β high β



Annealing schedule & convergence

Convergence to the global optimum is guaranteed if: $\beta_t \sim \ln t$

- $\Rightarrow$ robust optimization procedure
- $\Rightarrow$ but: $\beta_t \sim \ln t$ is **too slow** for practical problems
- $\Rightarrow$ therefore: $\beta_{t+1} = \tau \beta_t, \quad \tau \in [1.01, 1.30]$ (exponential annealing)
- $\Rightarrow$ additionally: the `State Update loop` has to be iterated often enough, e.g. $M = 500 - 2000$ ($\rightsquigarrow$ thermal equilibrium)

Examples

1. Finding the global optimum of cost function (with continuous variables)

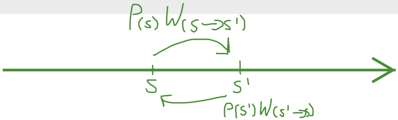
$\Rightarrow$ https://www.youtube.com/watch?v=iaq_Fpr4KZc

2. Solving Sudoku with Simulated Annealing

- initially fill columns randomly (without replacement)
- rows/3x3-boxes violate the Sudoku rules
- choose random column and two rows: switch the 2 numbers (stochastically)
- $s_i \in \{1, 2, \ldots, 9\} \implies (9!)^9 \geq 10^{50}$ states
- cost function $E_{(\underline{\mathbf{s}})}$ total number of doubles in all rows/boxes (normalized)
- multiple global optima and also local optima
- 1000 steps per `State Update loop`

$\Rightarrow$ https://www.youtube.com/watch?v=E8tkpzDne7I (from 2:19)

The Gibbs distribution



- for constant β: noisy state change via Markov process $\underline{\mathbf{s}}_{t'}$
- t': iteration count of the `State Update loop`
- $\Pi_{(\underline{\mathbf{s}},t')}$: probability distribution across states

$$\Pi_{(\underline{\mathbf{s}},t')} \rightarrow \underbrace{P_{(\underline{\mathbf{s}})}}_{\text{stationary distribution}} \quad \text{for } t' \rightarrow \infty \text{ (and constant } \beta)$$

$\rightarrow P(\underline{\mathbf{s}})$ can be calculated analytically!

Calculation of the stationary distribution

Assumption of *detailed balance:*

$$\underbrace{\text{probability of transition } \underline{\mathbf{s}}\to\underline{\mathbf{s}}'}_{P_{(\underline{\mathbf{s}})}\mathrm{W}_{(\underline{\mathbf{s}}\to\underline{\mathbf{s}}')}} = \underbrace{\text{probability of transition } \underline{\mathbf{s}}'\to\underline{\mathbf{s}}}_{P_{(\underline{\mathbf{s}}')}\mathrm{W}_{(\underline{\mathbf{s}}'\to\underline{\mathbf{s}})}}$$

$$\frac{P_{(\underline{\mathbf{s}})}}{P_{(\underline{\mathbf{s}}')}} = \frac{\mathrm{W}_{(\underline{\mathbf{s}}'\to\underline{\mathbf{s}})}}{\mathrm{W}_{(\underline{\mathbf{s}}\to\underline{\mathbf{s}}')}} = \frac{1+\exp\left\{\beta\left(\overbrace{E_{(\underline{\mathbf{s}})}-E_{(\underline{\mathbf{s}}')}}^{\Delta E}\right)\right\}}{1+\exp\left\{\beta\left(\underbrace{E_{(\underline{\mathbf{s}}')}-E_{(\underline{\mathbf{s}})}}_{-\Delta E}\right)\right\}} = \frac{1+\exp(\beta\Delta E)}{1+\exp(-\beta\Delta E)}$$

$$= \exp(\beta\Delta E)\frac{1+\exp(-\beta\Delta E)}{1+\exp(-\beta\Delta E)} = \exp(\beta\Delta E)$$

this condition is fulfilled for:

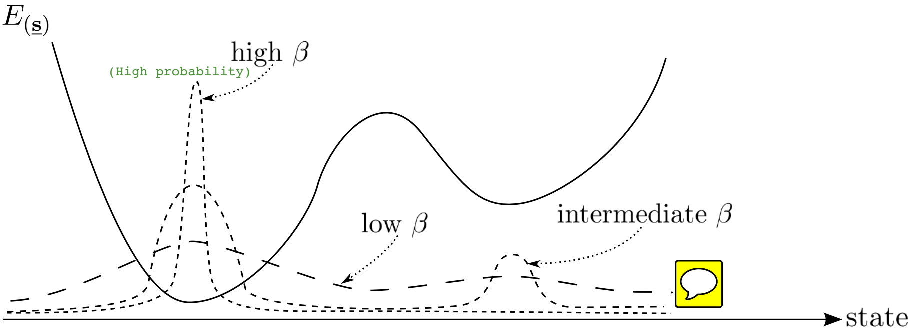
$$P_{(\underline{\mathbf{s}})} = \frac{1}{Z}\exp(-\beta E) \qquad \text{(Gibbs-Boltzmann-distribution)}$$

normalization constant / partition function: $Z = \sum_{\underline{\mathbf{s}}} \exp(-\beta E)$

Cost vs. probability distribution

$$P_{(\underline{\mathbf{s}})} = \frac{1}{Z} \exp(-\beta E)$$ (Gibbs-Boltzmann-distribution)

Based on local Distribution



$\beta \downarrow$: broad, "delocalized" distribution

$\beta \uparrow$: distribution localized around (global) minima

Mean-field annealing

Simulated Annealing

- $\rightarrow$ stochastic optimization: computationally expensive (sampling!)
- $\rightarrow$ stationary distribution $P_{(\underline{\mathbf{s}})}$ known (for each β_t), why not evaluate?
- $\rightarrow$ however: maxima of $P_{(\underline{\mathbf{s}})}$ equally hard to obtain as minima of $E_{(\underline{\mathbf{s}})}$
- $\rightarrow$ moments? for $\beta \rightarrow \infty$: $\langle \underline{\mathbf{s}} \rangle_P$ converges to $\underline{\mathbf{s}}^*$ of minimal cost ($P_{(\underline{\mathbf{s}})}$ singular)
- $\rightarrow$ but: moments of $P_{(\underline{\mathbf{s}})}$ can – in general – not be calculated analytically

Approximation by Mean-Field Annealing `Can we do better?`

- $\Rightarrow$ idea: approximate $P_{(\underline{\mathbf{s}})}$ by a computationally tractable distribution $Q_{(\underline{\mathbf{s}})}$
- $\Rightarrow$ this distribution is then used to calculate the first moment $\langle \underline{\mathbf{s}} \rangle_Q$
- $\Rightarrow$ the first moment is tracked during the annealing schedule β_t
- $\Rightarrow$ hope: $\langle \underline{\mathbf{s}} \rangle_Q \rightarrow \underline{\mathbf{s}}^*$ for $\beta_t \rightarrow \infty$

Factorizing distribution

$Q_{(\underline{s})} \sim \exp(-\beta \sum_k e_k s_k)$
$\sim \prod_k \exp(\beta e_k s_k)$

Distribution $Q_{(\underline{\mathrm{s}})}$ to approximate $P_{(\underline{\mathrm{s}})}$ Simple for Computing Moments

$$Q_{(\underline{\mathrm{s}})} \quad = \quad \frac{1}{Z_Q} \exp\{-\beta E_Q\} \quad = \quad \frac{1}{Z_Q} \exp\Big\{ -\beta \sum_k \underbrace{e_k}_{\text{parameters}} \mathrm{s}_k \Big\}$$

- Gibbs distribution with costs E_Q linear in the state variable $\underline{\mathrm{s}}_k$
- factorizing distribution $Q_{(\underline{\mathrm{s}})} = \Pi_k Q_k(s_k)$ with
  $Q_k(s_k) = \frac{1}{Z_{Q_k}} \exp(-\beta e_k \mathrm{s}_k)$
- $Q_{(\underline{\mathrm{s}})}$ factorizing $\iff$ s_k independent
  $\implies \langle \Pi_k s_k \rangle_Q = \Pi_k \langle s_k \rangle_Q \quad \binom{\text{moments}}{\text{factorize}}$
- $\big\langle s_k \big\rangle_Q = \dfrac{\sum\limits_{s_k} s_k \exp(-\beta e_k s_k)}{\sum\limits_{s_k} \exp(-\beta e_k s_k)}$

$\rightarrow$ family of distributions parametrized by the *mean fields* e_k

$\rightarrow$ determine e_k such that this approximation is as good as possible

Mean-field approximation

Quantities

$$P_{(\underline{\mathbf{s}})} = \frac{1}{Z_p} \exp(-\beta E_p)$$

true distribution

$$Q_{(\underline{\mathbf{s}})} = \frac{1}{Z_Q} \exp\Big(-\beta \overbrace{\sum_k e_k s_k}^{E_Q} \Big)$$

approximation: family of factorizing distributions

e_k : *mean fields*

parameters to be determined

Good approximation of P by Q

$\rightarrow$ minimization of the KL-divergence:

$$\mathrm{D_{KL}}(Q||P) = \sum_{\underline{\mathbf{s}}} Q_{(\underline{\mathbf{s}})} \ln \frac{Q_{(\underline{\mathbf{s}})}}{P_{(\underline{\mathbf{s}})}} \stackrel{!}{=} \min_{\underline{\mathbf{e}}}$$

Minimization of KL-divergence

k: Number of variables

$$\mathrm{D_{KL}}(Q||P) = \underbrace{\sum_{\underline{\mathbf{s}}}}_{\text{Sum of states}} Q_{(\underline{\mathbf{s}})} \ln \frac{Q_{(\underline{\mathbf{s}})}}{P_{(\underline{\mathbf{s}})}} \overset{!}{=} \min_{\underline{\mathbf{e}}} \qquad \begin{aligned} P_{(\underline{\mathbf{s}})} &= \frac{1}{Z_p}\exp(-\beta E_p) \\ Q_{(\underline{\mathbf{s}})} &= \frac{1}{Z_Q}\exp\Big(-\beta\sum_k e_k s_k\Big) \end{aligned}$$

Summe

$$\frac{\partial}{\partial e_l}\mathrm{D_{KL}} = \frac{\partial}{\partial e_l}\Big\{\beta\sum_{\underline{\mathbf{s}}} Q_{(\underline{\mathbf{s}})} E_p - \beta\sum_{\underline{\mathbf{s}}} Q_{(\underline{\mathbf{s}})} E_Q + \ln Z_p - \ln Z_Q\Big\}$$

Auch Summe (gesamt)

$$= \beta\frac{\partial}{\partial e_l}\big\langle E_p\big\rangle_Q \underbrace{-\beta\frac{\partial}{\partial e_l}\Big(\sum_{\underline{\mathbf{s}}} Q_{(\underline{\mathbf{s}})}\sum_k e_k s_k\Big)}_{-\beta\sum_k e_k\frac{\partial}{\partial e_l}\langle s_k\rangle_Q \cancel{-\beta\langle s_l\rangle_Q}} \underbrace{-\frac{1}{Z_Q}\sum_{\underline{\mathbf{s}}}\frac{\partial}{\partial e_l}\exp\Big(-\beta\sum_k e_k s_k\Big)}_{\cancel{+\beta\langle s_l\rangle_Q}}$$

$$= \boxed{\beta\frac{\partial}{\partial e_l}\big\langle E_p\big\rangle_Q - \beta\sum_k e_k\frac{\partial}{\partial e_l}\big\langle s_k\big\rangle_Q \overset{!}{=} 0,} \qquad l = 1,\ldots,N$$

Result

$$\frac{\partial}{\partial e_l}\langle E_p\rangle_Q - \sum_k e_k \frac{\partial}{\partial e_l}\langle s_k\rangle_Q = 0$$

s_k are independent under Q :

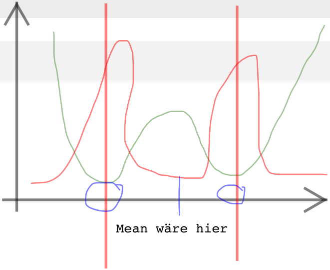
$$\left.\begin{aligned}
&\frac{\partial}{\partial e_l}\langle E_p\rangle_Q - e_l \frac{\partial}{\partial e_l}\langle s_l\rangle_Q = 0 \\
&\langle s_k\rangle_Q = \frac{\sum\limits_{s_k} s_k \exp(-\beta e_k s_k)}{\sum\limits_{s_k} \exp(-\beta e_k s_k)}
\end{aligned}\right\}$$ Zu lösende gleichungen

$\rightarrow$ coupled deterministic system of equations for $\{e_k\}$
$\rightarrow$ iterative solution procedure (usually no analytic result)

Mean-field annealing

Unterschied ist in der inner Loop

Algorithm



initialization: $\langle \underline{\mathbf{s}} \rangle_0, \beta_0$
BEGIN Annealing loop
  Repeat

- calculate mean-fields: $e_k, \quad k = 1, \ldots, N$
- calculate moments: $\left\langle s_k \right\rangle_Q, \quad k = 1, \ldots, N$

  Until $|e_k^{\text{old}} - e_k^{\text{new}}| < \varepsilon$ Accuracy
  increase β
END Annealing loop

$\Rightarrow$ inner loop: fixed-point iteration for the mean-fields e_k ($\rightsquigarrow$ EM-like)

$\Rightarrow$ deterministic (fast) rather than stochastic (slow) optimization method (given that mean-field equations can be easily evaluated, dep. on E_p)

$\Rightarrow$ moments $\left\langle s_k \right\rangle$ in general not from state space but $\left\langle s_k \right\rangle \to s_k^*$ for $\beta \to \infty$

Example (Ising model) – Setting and first Moments

Quadratic cost function $E(\underline{\mathbf{s}})$ with binary variables $s_k \in \mathcal{S} = \{+1, -1\}$,

$P_{(\underline{s})} \sim \exp(-\beta E_p)$

$$E_p(\underline{\mathbf{s}}) = -\frac{1}{2} \sum_{\substack{i=1,j=1 \\ i \neq j}}^{N} W_{ij} s_i s_j,$$

real symmetric matrix $\underline{\mathbf{W}}$, no self-coupling

$\rightarrow$ Expressions required for the mean-field algorithm can be calculated:

$$\langle s_k \rangle_Q = \frac{\sum_{s_k \in \mathcal{S}} s_k \exp(-\beta e_k s_k)}{\sum_{s_k \in \mathcal{S}} \exp(-\beta e_k s_k)} = \frac{(+1)\exp(-\beta e_k) + (-1)\exp(\beta e_k)}{\exp(-\beta e_k) + \exp(\beta e_k)}$$

$$= \boxed{\tanh(-\beta e_k)}$$

Example (Ising model) – Mean-fields

$$
\begin{aligned}
0 &= \frac{\partial}{\partial e_k}\langle E_p\rangle_Q - e_k \frac{\partial}{\partial e_k}\langle s_k\rangle_Q \\
&= \frac{\partial}{\partial e_k}\left\langle -\frac{1}{2}\sum_{\substack{i=1,j=1\\ i\neq j}}^{N} W_{ij} s_i s_j \right\rangle_Q - e_k \frac{\partial}{\partial e_k}\langle s_k\rangle_Q \\
&= -\frac{1}{2}\frac{\partial}{\partial e_k}\sum_{\substack{i=1,j=1\\ i\neq j}}^{N} W_{ij}\langle s_i\rangle_Q \langle s_j\rangle_Q - e_k \frac{\partial}{\partial e_k}\langle s_k\rangle_Q \\
&= -\sum_{\substack{i=1\\ i\neq k}}^{N} W_{ik}\langle s_i\rangle_Q \frac{\partial}{\partial e_k}\langle s_k\rangle_Q - e_k \frac{\partial}{\partial e_k}\langle s_k\rangle_Q
\end{aligned}
$$

$$\frac{\partial}{\partial e_k}\langle s_k\rangle_Q \underbrace{\left\{-\sum W_{ik}\langle s_i\rangle_Q - e_k\right\}}_{0} = 0$$

$$\Longrightarrow \boxed{e_k = -\sum_{\substack{i=1\\ i\neq k}}^{N} W_{ik}\langle s_i\rangle_Q}$$

(will be applied in exercise sheet 9)

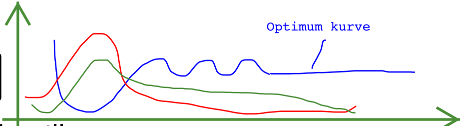
Example (Ising model) – Fixed point iteration

Inner loop in mean-field annealing algorithm:

`Repeat`

- calculate mean-fields: $e_k = -\sum_{\substack{i=1 \\ i \neq k}}^{N} W_{ik} \langle s_i \rangle_Q, \quad k = 1, \ldots, N$
- calculate moments: $\langle s_k \rangle_Q = \tanh(-\beta e_k), \quad k = 1, \ldots, N$

`Until` $|e_k^{\text{old}} - e_k^{\text{new}}| < \varepsilon$



$\rightsquigarrow$ fixed-point iteration for mean-fields will converge