Применение квантового отжига к задачам машинного обучения

Игорь Побойко

18 апреля 2018 г.

Семинар “Вычислительные среды”, НИУ ВШЭ
По мотивам arXiv:1706.08470v2,
Carlo Baldassi and Riccardo Zecchina

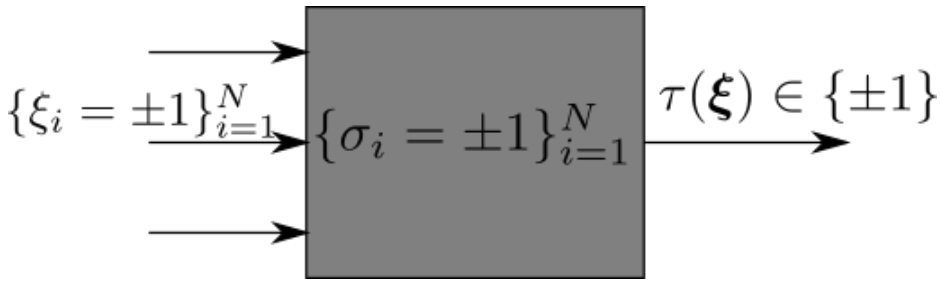
План

- Введение
  - Исследуемая задача, её основные свойства
  - Симулированный (классический) отжиг
  - Квантовый отжиг и квантовый Монте-Карло
- Основные результаты работы
- Интерпретация

Binary perceptron

Бинарный перцептрон — "нейрон":

$$\tau(\boldsymbol{\xi}) = \mathrm{sign}(\boldsymbol{\xi} \cdot \boldsymbol{\sigma})$$



- Задача: имеется αN «входов» $\boldsymbol{\xi}^\mu$ и соответствующих им «выходов» τ^μ. Хотим найти состояние $\boldsymbol{\sigma}$, которое им всем удовлетворит.
- Соответствующая задача оптимизации: «энергия» считает количество ошибок:

$$E(\boldsymbol{\sigma}) = \sum_{\mu=1}^{\alpha N} \theta(-\Delta_\mu), \quad \Delta_\mu = \frac{\tau^\mu(\boldsymbol{\xi}^\mu \cdot \boldsymbol{\sigma})}{\sqrt{N}}$$

  и мы хотим эту энергию минимизировать (в идеале — до нуля)
- Задача будет изучаться в термодинамическом пределе ($N \gg 1$) на случайных наборах $\boldsymbol{\xi}^\mu$ и τ^μ.

Fun facts

- В задаче имеется фазовый переход при $\alpha_c \approx 0.83$.
- При $\alpha < \alpha_c$, решений экспоненциально много; при $\alpha > \alpha_c$, их нет.

M. Mezard, J. Phys. A: Math. Gen. **22** (1989)

- Задача «стекольного» типа: много локальных метастабильных минимумов, разделённых экспоненциально большими энергетическими барьерами
- Решения группируются в кластеры с экспоненциально большой их плотностью

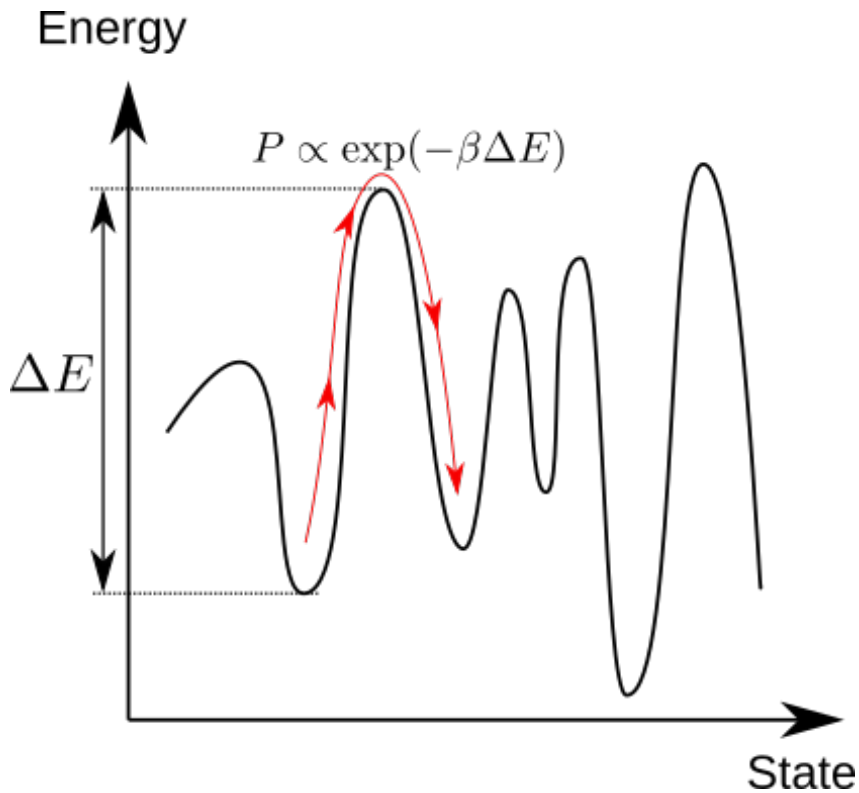
Carlo Baldassi et al., Phys. Rev. Lett. **115**, 128101 (2015)

Симулированный отжиг (Simulated annealing, SA)

- Рассмотрим Гиббсовский ансамбль:

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp(-\beta E(\boldsymbol{\sigma}))$$

- Будем сэмплировать это распределение Монте-Карло (Metropolis, Heat bath, etc.)
- Контроллируемо увеличиваем β, добиваясь термализации
- В пределе $\beta \to \infty$, энергия минимизирована!



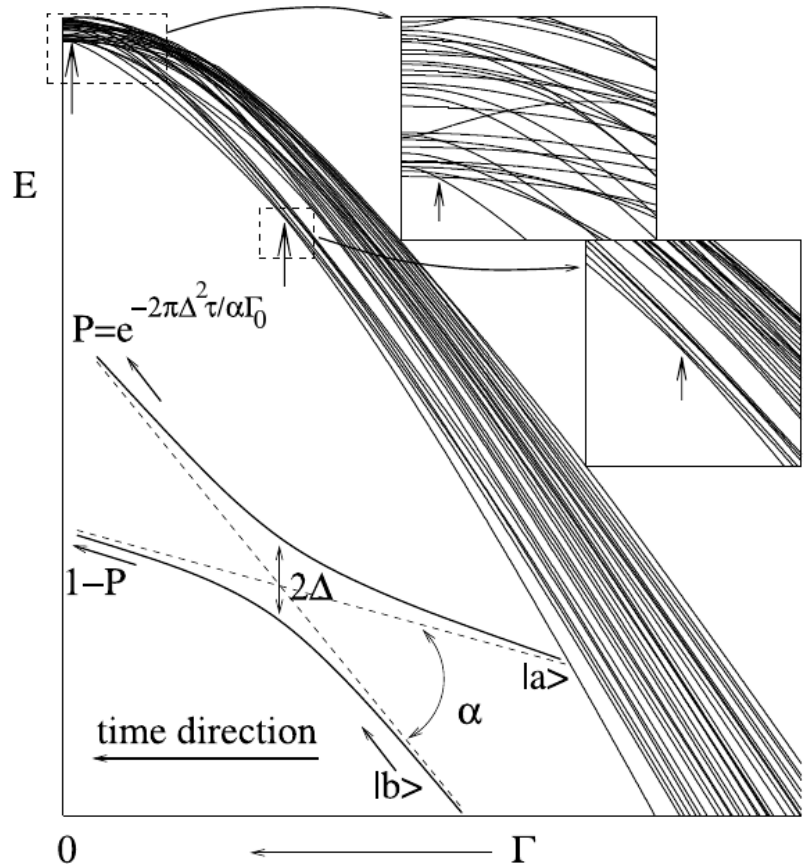
Много метастабильных состояний — алгоритм "застряёт"!
Требуется экспоненциально большое время, чтобы достигнуть оптимума.

Квантовый отжиг (Quantum Annealing, QA)

- Рассмотрим гамильтониан:

$$\hat{H} = E(\hat{\boldsymbol{\sigma}}^z) - \Gamma(t) \sum_{i=1}^{N} \hat{\sigma}_i^x$$

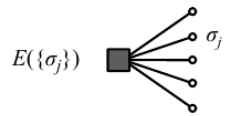
- Пусть $\Gamma(0) \gg 1$, тогда основное состояние $|\psi\rangle = \prod_{i=1}^{N} \frac{|\uparrow\rangle_i + |\downarrow\rangle_i}{\sqrt{2}}$, и щель $\Delta \sim 2\Gamma$.
- Адиабатическая теорема: если $\Gamma(t)$ меняется плавно, то система останется в основном состоянии!
- Уменьшим до нуля, $\Gamma(T) = 0$ — PROFIT.



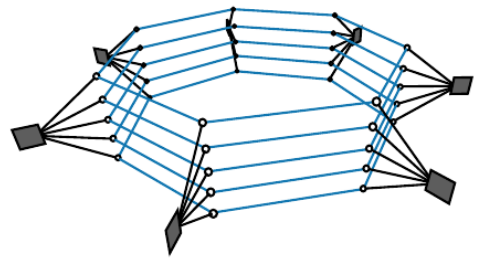
Насколько медленно? По сравнению с типичными щелями в спектре! Щели определяются туннельным перекрытием: $\Delta \propto \exp(-\#d(\boldsymbol{\sigma}, \boldsymbol{\sigma}^*))$ — они тоже экспоненциально малы, но всё не так уж плохо...

Симулированный квантовый отжиг (Simulated Quantum Annealing, SQA)

$$\hat{H} = E(\hat{\boldsymbol{\sigma}}^z) - \Gamma \sum_{i=1}^{N} \hat{\sigma}_i^x, \quad Z = \mathrm{Tr}\exp(-\beta \hat{H})$$



Квантовый МК $\equiv$ классический МК на реплецированной системе:



$$H_{\mathsf{eff}}[\boldsymbol{\sigma}] = \frac{1}{y}\sum_{a=1}^{y} E(\boldsymbol{\sigma}^a) - \frac{\gamma}{\beta}\sum_{a=1}^{y}\sum_{j=1}^{N} \sigma_j^a \sigma_j^{a+1} - \frac{NK}{\beta},$$

$$\gamma = \frac{1}{2}\ln\coth\frac{\beta\Gamma}{y}, \quad K = \frac{y}{2}\ln\left(\frac{1}{2}\sinh\frac{2\beta\Gamma}{y}\right)$$

- Хотим: $y \to \infty$, $\beta \to 0$ (ground state)
- Аналогично классическому отжигу, плавно понижаем Γ до 0
- Такой протокол примерно эквивалентен QA.

Sergei V. Isakov et al., arXiv:1510.08057v1 (2015)

Техническое отступление (о QMC)

Статсумма:

$$Z = \mathrm{Tr}(e^{-\beta\hat{H}}) = \mathrm{Tr}\left(\underbrace{e^{-\frac{\beta}{y}E(\hat{\boldsymbol{\sigma}}^z)}e^{\frac{\beta}{y}\Gamma\hat{\sigma}_j^x}\ldots e^{-\frac{\beta}{y}E(\hat{\boldsymbol{\sigma}}^z)}e^{\frac{\beta}{y}\Gamma\sum_j\hat{\sigma}_j^x}}_{y\ \textbf{terms}}\right) =$$

$$= \sum_{\{\sigma_j\}} e^{-\frac{\beta}{y}\sum_{a=1}^{y}E(\boldsymbol{\sigma}_a)}\left\langle\boldsymbol{\sigma}^1\right|e^{\frac{\beta\Gamma}{y}\sum_j\hat{\sigma}_j^x}\left|\boldsymbol{\sigma}^2\right\rangle\ldots\left\langle\boldsymbol{\sigma}^y\right|e^{\frac{\beta\Gamma}{y}\sum_j\hat{\sigma}_j^x}\left|\boldsymbol{\sigma}^1\right\rangle$$

Матричные элементы:

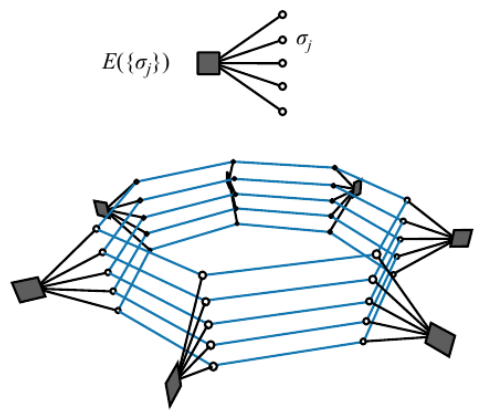
$$e^{\frac{\beta\Gamma}{y}\hat{\sigma}^x} = \cosh\frac{\beta\Gamma}{y} + \hat{\sigma}^x\sinh\frac{\beta\Gamma}{y} = \begin{pmatrix}\cosh\frac{\beta\Gamma}{y} & \sinh\frac{\beta\Gamma}{y}\\ \sinh\frac{\beta\Gamma}{y} & \cosh\frac{\beta\Gamma}{y}\end{pmatrix} =$$

$$= \exp\left(\frac{1}{2}\ln\left(\frac{1}{2}\sinh\frac{2\beta\Gamma}{y}\right) + \frac{1}{2}\sigma_1\sigma_2\ln\coth\frac{\beta\Gamma}{y}\right)$$

Получается классическое распределение Гиббса по y “репликам” с “эффективным гамильтонианом”:

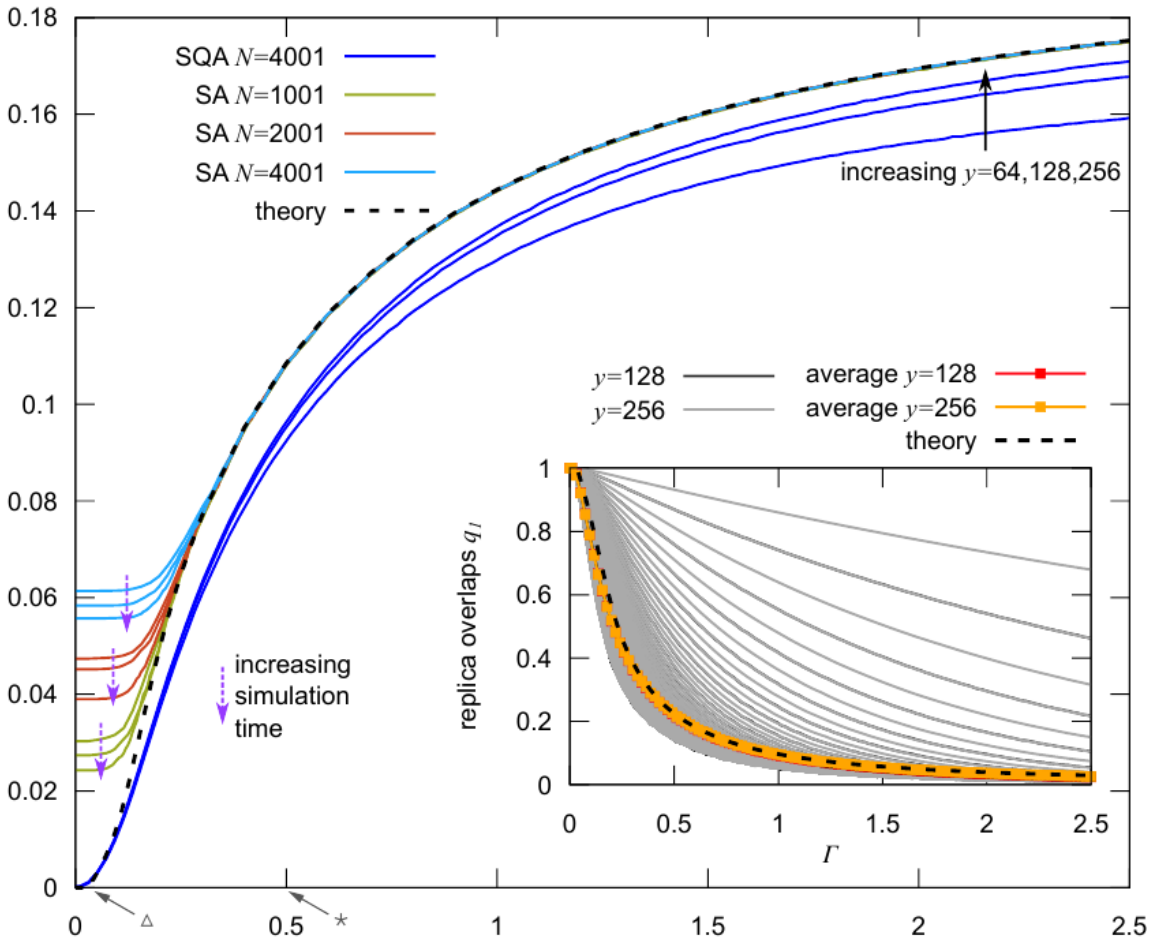
$$H_{\textbf{eff}}[\boldsymbol{\sigma}] = \frac{1}{y}\sum_{a=1}^{y}E(\boldsymbol{\sigma}^a) - \frac{\gamma}{\beta}\sum_{a=1}^{y}\sum_{j=1}^{N}\sigma_j^a\sigma_j^{a+1} - \frac{NK}{\beta},$$

$$\gamma = \frac{1}{2}\ln\coth\frac{\beta\Gamma}{y}, \quad K = \frac{y}{2}\ln\left(\frac{1}{2}\sinh\frac{2\beta\Gamma}{y}\right)$$



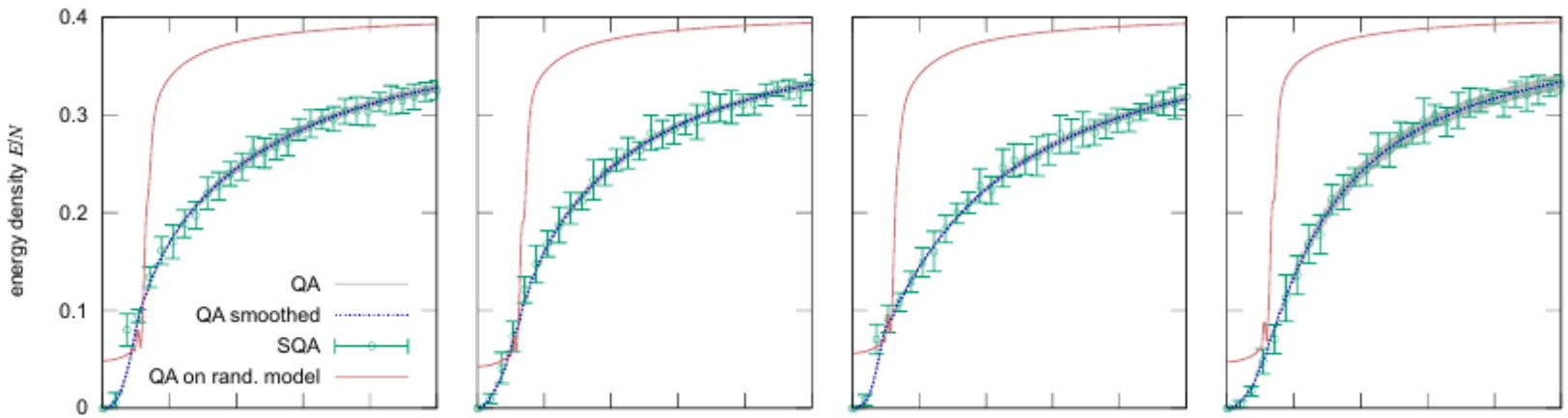
Результаты (SQA vs SA)

- Протокол: 30τ линейных шагов по Γ, $\alpha = 0.4$
- SA: $\beta = \beta(\Gamma)$, полное число итераций — $\tau N \cdot 10^4$ $(\tau = 4, 8, 16)$
- SQA: $\beta = 20$, полное число итераций — $\tau N y \cdot 10^4$ $(\tau = 1, 2, 4)$



Результаты: randomization (QA, Exact diagonalization)

- Численное решение УШ $i\frac{\partial|\psi(t)\rangle}{\partial t} = \hat{H}(t)\,|\psi(t)\rangle$, при $\Gamma(t) = 5 \cdot (1 - t/1000)$, $N = 21$ (разреженная матрица размера $2^{21} \sim 2 \cdot 10^6$)
- Выборка образцов: много решений ($\geq N$), и SA плох.
- Benchmark: те же образцы, случайно перемешиваются энергии (randomized samples). Термодинамика такая же, а энергетический ландшафт другой.



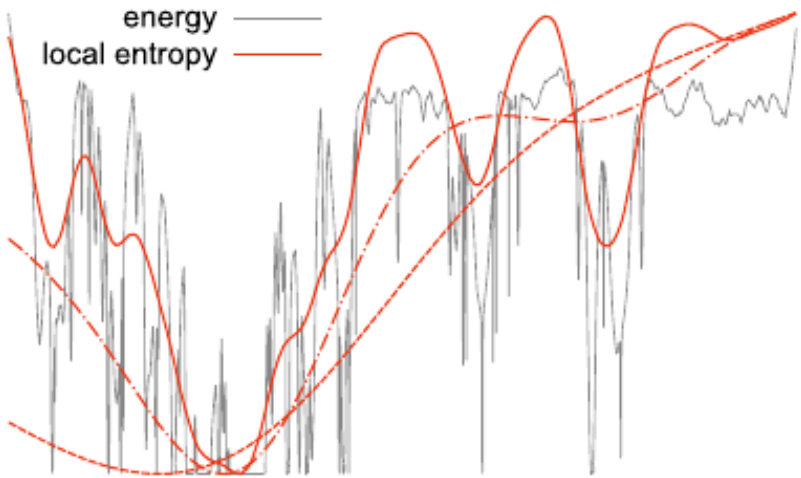

And Now For Something
Completely Different

Robust ensemble (1)

- Мотивация: есть классические эффективные эвристики. Эмпирический факт: они находят области с большим количеством решений в окрестности.
- "Локальная свободная энтропия" и RE:

$$\Phi(\boldsymbol{\sigma}^*) = \ln \sum_{\{\boldsymbol{\sigma}\}} e^{-\beta E(\boldsymbol{\sigma}) - \gamma d(\boldsymbol{\sigma}, \boldsymbol{\sigma}^*)}, \quad P_{RE}(\boldsymbol{\sigma}^*) \propto e^{y\Phi(\boldsymbol{\sigma}^*)}$$

- При $\beta \to \infty$, $\Phi(\boldsymbol{\sigma})$ "считает" количество минимумов рядом с $\boldsymbol{\sigma}$. При $y \gg 1$, эти конфигурации имеют большой стат. вес
- $\gamma \to \infty$ соответствует исходному ансамблю



Proceedings of the National Academy of Sciences 113, E7655 (2016)

Robust ensemble (2)

$$\Phi(\boldsymbol{\sigma}^*) = \ln \sum_{\{\boldsymbol{\sigma}\}} e^{-\beta E(\boldsymbol{\sigma}) - \gamma d(\boldsymbol{\sigma}, \boldsymbol{\sigma}^*)}, \quad P_{RE}(\boldsymbol{\sigma}^*) \propto e^{y\Phi(\boldsymbol{\sigma}^*)}$$

Позволяет построить классические эффективные алгоритмы (Entropy-SGD)

Model	Entropy-SGD		SGD / Adam	
	Error (%) / Perplexity	Epochs	Error (%) / Perplexity	Epochs
mnistfc	1.37 ± 0.03	120	1.39 ± 0.03	66
LeNet	0.5 ± 0.01	80	0.51 ± 0.01	100
All-CNN-BN	7.81 ± 0.09	160	7.71 ± 0.19	180
PTB-LSTM	77.656 ± 0.171	25	78.6 ± 0.26	55
char-LSTM	1.217 ± 0.005	25	1.226 ± 0.01	40

Pratik Chaudhari et al., arXiv:1611.01838v5 (2017)

“Связь” RE и SQA

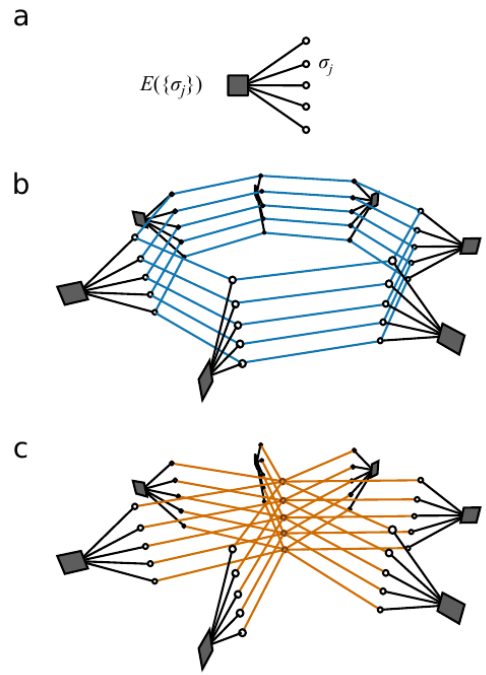
- SQA: y реплик $\boldsymbol{\sigma}^a$:

$$P_{\text{SQA}} \propto e^{-\frac{\beta}{y}\sum_{a=1}^{y} E(\boldsymbol{\sigma}^a) - \gamma \sum_{a=1}^{y} \boldsymbol{\sigma}^a \boldsymbol{\sigma}^{a+1}}$$

- RE: исходная конфигурация $\boldsymbol{\sigma}^*$
плюс y реплик $\boldsymbol{\sigma}^a$:

$$P_{\text{RE}} \propto e^{-\beta \sum_{y=1}^{a} E(\boldsymbol{\sigma}^a) - \gamma \sum_{a=1}^{y} d(\boldsymbol{\sigma}^*, \boldsymbol{\sigma}^a)}$$

- При $\gamma \to \infty$ и $\Gamma \to 0$, всё совпадает
с исходным ансамблем



Локальные энергетические профили

- Belief Propagation — составляются Bethe-Pierls-like уравнения на локальные вероятности переменных, которые итеративно решаются.

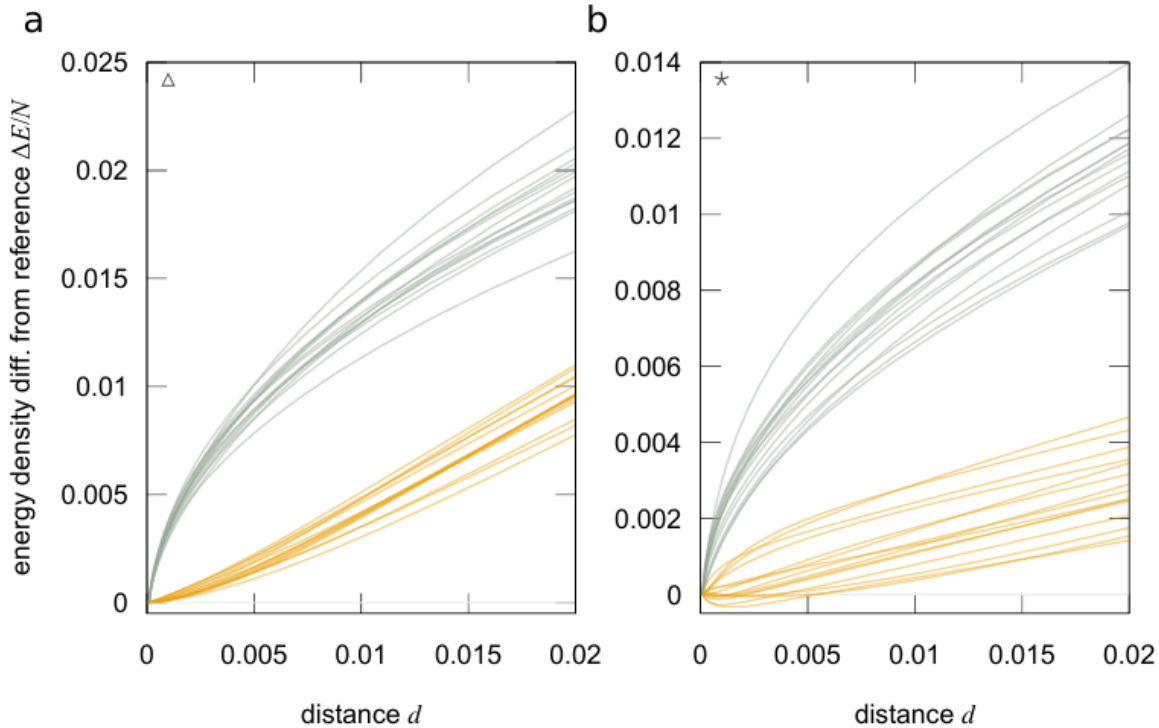


Рис.: Оранжевые линии — QA, серые — SA (Δ: $\Gamma \approx 0$, $*$: $\Gamma \approx 0.5$)

Заключение

Messages to take home:

- В задачах ML имеется особая структура локальных минимумов: широкие долины
- Квантовый отжиг (предположительно) любит такие задачи

Спасибо за внимание!