Sampling based optimization

Richard Combes (rcombes@kth.se)
Jie Lu
Alexandre Proutière

FEL 3310: Distributed optimization

The original problem: Maxwell-Boltzman statistics

- Original problem: calculation of Maxwell-Boltzman statistics
- Model for non-interacting particles (i.e perfect gas).
- Thermodynamical system, state s, state space $\mathcal{S}$ finite.
- Potential energy of a state $E(s)$, temperature $T > 0$, b Boltzmann constant.
- At thermodynamical equilibrium, the system state follows the Boltzmann distribution:

$$p(s) = \frac{\exp(-\frac{E(s)}{bT})}{\sum_{s' \in \mathcal{S}} \exp(-\frac{E(s')}{bT})}$$

- Problem: $|\mathcal{S}|$ large, $\sum_{s' \in \mathcal{S}} \exp(-\frac{E(s')}{bT})$ impossible to calculate directly.

The first MCMC method: Metropolis-Hastings

- Solution (Metropolis, 1953): define a Markov chain $\{X_n\}$ which admits p as a stationary distribution
- Result obtained by averaging

$$\frac{1}{t}\sum_{n=1}^{t} f(X_n) \to_{t\to+\infty} \sum_{s\in\mathcal{S}} p(s)f(s) \text{ a.s.}$$

- Define $N(s) \subset \mathcal{S}$ neigbours of s. Symmetry: $s' \in N(s)$ iff $s \in N(s')$.
- Metropolis-Hastings algorithm:

$$\begin{aligned} X_0 &\in \mathcal{S} \\ Y_n &\sim \text{Uniform}(N(X_n)) \\ X_{n+1} = Y_n &\text{ with proba } \min(e^{-\frac{E(Y_n)-E(X_n)}{bT}}, 1) \\ X_{n+1} = X_n &\text{ otherwise.} \end{aligned}$$

The first MCMC method: Metropolis-Hastings

- Transition probability, $s' \in N(s)$:

$$P(s,s') = \frac{\min(e^{-\frac{E(s')-E(s)}{bT}}, 1)}{|N(s)|}.$$

- X_n reversible Markov chain with stationary distribution p (detailed balance holds):

$$p(s)P(s,s') = p(s')P(s',s),$$

- If N is large: low probability of changing, if N is small, takes time to go through the state space.

MCMC: sampling a distribution known up to a constant

- General problem: distribution $p(.)$ known up to a constant on a high dimensional space, how to sample from p ?
- Ingredients: $Q(.,.)$ (symmetrical) proposal distribution, $R(.,.)$ acceptance probability
- Basic algorithm:

$$\begin{aligned} X_0 &\in \mathcal{S} \\ Y_n &\sim Q(X_n,.) \\ X_{n+1} &= Y_n \text{ with probability } R(X_n, Y_n) \\ X_{n+1} &= X_n \text{ with probability } 1 - R(X_n, Y_n). \end{aligned}$$

- Detailed balance equations impose:

$$R(s, s') = \begin{cases} 1 & \text{if } p(s') \geq p(s) \\ \frac{p(s')}{p(s)} & \text{otherwise.} \end{cases}$$

MCMC: the impact of mixing

- The sequence generally moves towards regions of high probability
- Advantage over rejection sampling: the proposal distribution is a function of the samples
- Disadvantage: samples are correlated
- Efficiency measured by the mixing time: successive samples should be *as de-correlated as possible*.
- Choice of Q is *critical*:
  - large jumps: most states have very low probability, acceptance probability is low, so the chain stays static most of the time
  - small jumps: the chain takes a lot of time to go through the state space.
- Choosing Q is not straightforward.

Sampling per component: Gibbs Sampling

- Going back to the first example, consider K particles each with 2 possible states.
- State space, $\mathcal{S} = [0, 1]^K$, state $s = (s_1, \ldots, s_K)$.
- k-th particle , state: $s = (s_k, s_{-k})$,
- Joint distribution p is complex, however $p(s_k|s_{-k})$ is very simple (Bernoulli distribution):

$$p(s_k = 0|s_{-k}) = \frac{e^{-\frac{E(0,s_{-k})}{bT}}}{e^{-\frac{E(0,s_{-k})}{bT}} + e^{-\frac{E(1,s_{-k})}{bT}}}.$$

- Idea of Gibbs sampling (Geman , 1984): at each step, change the state of at most 1 particle.

Sampling per component: Gibbs Sampling

- Gibbs sampler: a sampling method for p (known up to a constant), when conditionals $p(x_k|x_{-k})$ are easy to calculate
- At each step, change a component selected at random.

$$\begin{aligned}
X_0 &\in \mathcal{S} \\
k(n) &\sim \text{Uniform}(\{1, \ldots, K\}) \\
Y_n &\sim p(\,.\,|X_{n,-k(n)}) \\
X_{n+1,k(n)} &= Y_n \\
X_{n+1,k} &= X_{n,k} \text{ if } k \neq k(n)
\end{aligned}$$

- No rejection in Gibbs sampling.
- Lends itself to distributed implementation.
- Blocked Gibbs sampler: same method with blocks of variables

Simulated annealing

- $\mathcal{S}$ finite set, cost function $V : \mathcal{S} \to \mathbb{R}^+$
- Goal: minimize V, set of minima $H = \{\arg\max_s V(s)\}$.
- Boltzmann distribution:

$$p(s,T) = \frac{\exp(-\frac{V(s)}{T})}{\sum_{s' \in \mathcal{S}} \exp(-\frac{V(s')}{T})}$$

- At low temperatures, $p(.,T)$ is concentrated on H, $p(H,T) \to 1$, $T \to 0^+$.
- Intuition: sample from p using MCMC while decreasing T
- Cooling schedule: $T \to 0$ slowly enough so that $X_n \to_{n\to\infty} H$ a.s.
- Annealing principle, analogy with solid state physics: first heat then slowly cool a metal to improve its crystalline structure. Minimal potential = perfect crystal.

Cooling schedules

- Main question: which cooling schedules ensure convergence ?
- Here we study a simple case: the schedule is constant by parts.
- Step $m \in \mathbb{N}$ of duration α_m, $t_m = \sum_{m'<m} \alpha_{m'}$.
- Cooling schedule: $T = T_m$, $t \in [t_m, t_m + \alpha_m]$
- Intuition: if α_m is large with respect to the mixing time at temperature T_m, $X_{t_{m+1}}$ should follow $p(., T_m)$

A convergence theorem

Define: $\delta = \min_{s \notin H} V(s)$, $V_\infty = \max_{s \in \mathcal{S}} V(s)$.

Theorem

There exists $a_0 > 0$ such that by choosing $T_m = \frac{\delta}{\log(m)}$, $\alpha_m = m^a$, $a \geq a_0$, the simulated annealing converges:

$$X_{t_m} \to_{m \to \infty} H, \ a.s\ .$$

A convergence theorem: proof

Lemma

There exists a positive sequence $\{\beta_m\}$ such that if for all m, $\alpha_m \geq \beta_m$, and $T_m = \frac{\delta}{\log(m)}$, then:

$$X_{t_m} \to_{m \to \infty} H, \; a.s\ .$$

Mixing time of reversible Markov chains

- Ergodic flow between subsets $\mathcal{S}_1, \mathcal{S}_2$:

$$K(\mathcal{S}_1, \mathcal{S}_2) = \sum_{s_1 \in \mathcal{S}_1} \sum_{s_2 \in \mathcal{S}_2} p(s_1) P(s_1, s_2),$$

- Conductance of the chain

$$\Phi = \min_{\mathcal{S}' \subset \mathcal{S}, p(\mathcal{S}') \leq 1/2} \frac{K(\mathcal{S}', \mathcal{S} \setminus \mathcal{S}')}{p(\mathcal{S}')}.$$

- Mixing time:

$$\tau(\epsilon) = \min\{n : \sup_s |\mathbb{P}(X_n = s) - p(s)| \leq \epsilon\}. \qquad (1)$$

Theorem

With the above definitions, and $p^* = \min_s p(s)$, we have:

$$\tau(\epsilon) \leq \frac{2}{\Phi^2}(\log(1/p^*) + \log(1/\epsilon)).$$

Payoff-based learning

- Principle: N independent agents with finite action sets want to minimize a function without any information exchange
- Agent i chooses $a_i \in \mathcal{A}_i$ and observes payoff $U_i(a_1, \ldots, a_N) \in [0, 1)$
- Goal: maximize $U(a) = \sum_{i=1}^{N} U_i(a)$, $H = \arg\max_a U(a)$
- "Payoff-based learning": agents do not observe the payoffs or actions of the other players.
- Assumption: agents cannot be separated in 2 disjoint subsets that do not interact.

Payoff based learning: a sampling method

- Sampling approach proposed by (Peyton-Young, 2012): design a Markov chain whose stationary distribution is concentrated on H
- State of agent i: $\overline{a}_i \in \mathcal{A}_i$ benchmark action, $\overline{u}_i \in [0, 1)$ benchmark payoff, “mood”$m_i \in \{C, D\}$ (“Content’ , “Discontent’)
- Experimentation rate $\epsilon > 0$, constant $c > N$.

Payoff based learning: update mechanism

If *i* is content:

- Choose action a_i:

$$\mathbb{P}[a_i = a] = \begin{cases} \epsilon^c/(|\mathcal{A}_i| - 1) & a \neq \overline{a}_i \\ 1 - \epsilon^c & a = \overline{a}_i \end{cases}$$

- Observe resulting u_i:
  - If $(a_i, u_i) = (\overline{a}_i, \overline{u}_i)$, *i* stays content
  - If $(a_i, u_i) \neq (\overline{a}_i, \overline{u}_i)$: *i* becomes discontent with probability $1 - \epsilon^{1-u_i}$.
- Benchmark actions are updated $(a_i, u_i) \leftarrow (\overline{a}_i, \overline{u}_i)$

If *i* is discontent:

- Choose action a_i:

$$\mathbb{P}[a_i = a] = 1/|\mathcal{A}_i| \ , \ a \in \mathcal{A}_i$$

- Observe resulting u_i, and become content with probability ϵ^{1-u_i}
- Benchmark actions are updated $(a_i, u_i) \leftarrow (\overline{a}_i, \overline{u}_i)$

Rationale of Peyton-Young's method

- Experiment (a lot) until content: When an agent is discontent, he plays an action at random, and becomes content only if he has chosen an action yielding high reward
- Do not change if content: An agent that is content remembers the (action,reward) that caused him to become content, so he keeps playing that same action with overwhelming probability
- Become discontent when others change: (change detection mechanism) whenever a content agent detects a change in reward he becomes discontent, because it indicates that another agent has deviated
- Experiment (a little) if content: Occasionally a content agent experiments (mandatory to avoid local minima)

A concentration result

Theorem

Consider the (irreducible) Markov chain $(\overline{u}_i, \overline{a}_i, m_i)_i$, denote by $p(., \epsilon)$ its stationary distribution. Define

$$H = \{(\overline{u}, \overline{a}, m) : \overline{u}_i = U_i(\overline{a}), \overline{a} \in H, m_i = C \ , \ \forall i\}.$$

Then H is the only stochastically stable set so that:

$$p(H, \epsilon) \to 1, \epsilon \to 0^+.$$

Resistance trees

- Main difficulty: the chain is *not reversible* .
- The proof is based on the theory of stochastic potential for perturbed Markov chains (Peyton-Young 1993).
- Perturbed Markov Chain: $P(s, s', \epsilon) \sim \epsilon^{r(s,s')}$, $\epsilon \to 0$
- $E_1, \dots, E_M$ recurrence classes of $P(., ., 0)$
- $r(s, s')$ resistance of link (s, s')
- Path from s to s', $\xi = (s = s_1, \dots, s_b = s')$, resistance is additive on paths:

$$r(\xi) = r(s_1, s_2) + \cdots + r(s_{b-1}, b_a).$$

Resistance trees

- Potential: $\rho_{i,j} = \min_\xi r(\xi)$; minimum is taken on all paths from $E_i \to E_j$.
- Define $\mathcal{G}$ weighted graph with vertices $\{1, \ldots, M\}$ and weights $(\rho_{i,j})_{1 \leq i,j \leq M}$.
- Fix i, consider a directed tree $\mathcal{T}$ on $\mathcal{G}$ which contains exactly one path from j to i (for all $j \neq i$).
- The stochastic potential of class i is the minimum of $\sum_{(i,j) \in T} \rho_{i,j}$, where the minimum is taken over all possible trees $\mathcal{T}$.

Theorem

The only stochastically stable recurrence classes $E_1, \ldots, E_M$ are the ones with minimum stochastic potential.

Some good reading

- Metropolis-Hastings: Metropolis, “Equations of State Calculations by Fast Computing Machines”
- MCMC: Andrieu, “An Introduction to MCMC for Machine Learning”
- Gibbs sampling: Geman, “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images”
- Markov chain mixing time: Levin, “Markov Chains And Mixing Times”
- Simulated Annealing: Hajek, “Cooling Schedules for Optimal Annealing ”
- Payoff-based learning: Peyton-Young, “The evolution of conventions”