Lecture 6: Monte Carlo Optimization

Tianpei Xie

Sep. 28th., 2022

Contents

1 Introduction

- Similar to the problem of integration, differences between the numerical approach and the simulation approach to the problem

$$\max_{\boldsymbol{x}\in\mathcal{X}} h(\boldsymbol{x}) \tag{1}$$

  lie in the treatment of the function h.

  - In approaching an optimization problem using deterministic numerical methods, the ***analytical properties*** of the target function (*convexity*, *boundedness*, *smoothness*) are often paramount.
  - For the simulation approach, we are more concerned with h from a ***probabilistic*** (rather than analytical) point of view.

- We want to distinguish between two approaches to ***Monte Carlo optimization*** [Robert and Casella, 1999].

  - The first is an ***exploratory approach***, in which the goal is to optimize the function h by *describing its entire range*. The *actual properties* of the function play a *lesser role* here, with the Monte Carlo aspect more closely tied to the ***exploration of the entire space*** $\mathcal{X}$, even though, for instance, the slope of h can be used to speed up the exploration.
  - The second approach is based on a ***probabilistic approximation*** of the *objective function* h and is somewhat of a *preliminary step* to the actual optimization. Here, the Monte Carlo aspect ***exploits the probabilistic properties of the function*** h to come up with an acceptable approximation and is less concerned with exploring $\mathcal{X}$. For instance, *Missing data methods*, such as the *EM algorithm*, are closely related to tied to this idea.

2 Stochastic Explorations

- There are a number of cases where the exploration method is particularly well suited. For instance, when $\mathcal{X}$ is bounded, we can use grid search, there the function h is approximated by $\boldsymbol{h} = [h(x_i)]_i$. Distributions other than the uniform, which can possibly be related to h, may then do better. In particular, in setups where the likelihood function is extremely costly to compute, the number of evaluations of the function h is best kept to a minimum.
- The second direction is to relates h to a probability distribution. For instance, $h > 0$ and $\int_{\mathcal{X}} h(\boldsymbol{x})d\boldsymbol{x} < \infty$, the resolution of (1) amounts to finding the ***modes*** of the density h. More generally, if these conditions are not satisfied, then we may be able to transform the function $h(\boldsymbol{x})$ into another function $H(\boldsymbol{x})$ that satisfies the condition of positivity with finite integral. For instance $H(\boldsymbol{x}) :\propto \exp(h(\boldsymbol{x})/T)$ or $H(\boldsymbol{x}) :\propto \sigma(h(\boldsymbol{x})/T)$ where $\sigma(x) = 1/(1+\exp(-x))$. We can choose T to accelerate convergence or to avoid local maxima (*Simulated Annealing*).

2.1 Gradient Methods

- The gradient method is a *deterministic* numerical approach to the problem (1). It produces a sequence $(\boldsymbol{x}_t)$ that converges to the exact solution of (1), $\boldsymbol{x}^*$, when the domain $\mathcal{X} \subseteq \mathbb{R}^d$ and the function $(-h)$ are both *convex*. The sequence $(\boldsymbol{x}_t)$ is constructed in a ***recursive*** manner

through

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \alpha_t \nabla h(\boldsymbol{x}_t), \quad \alpha_t > 0.$$

- In more general setups (that is, when the function or the space is less regular), equation above can be modified by stochastic perturbations to again achieve convergence. One of these stochastic modifications is to choose a second sequence (β_t) to define the chain $(\boldsymbol{x}_t)$ by

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \frac{\alpha_t}{2\beta_t} \Delta h(\boldsymbol{x}_t, \beta_t \boldsymbol{\xi}_t) \boldsymbol{\xi}_t, \quad \alpha_t > 0. \tag{2}$$

$$\text{where } \Delta h(\boldsymbol{x}_t, \beta_t \boldsymbol{\xi}_t) = (h(\boldsymbol{x}_t + \beta_t \boldsymbol{\xi}_t) - h(\boldsymbol{x}_t - \beta_t \boldsymbol{\xi}_t)) \approx 2\,\beta_t \left\|\boldsymbol{\xi}_t\right\|_2 \nabla h(\boldsymbol{x}_t),$$

and the variables $\boldsymbol{\xi}_t$ are uniformly distributed on the unit sphere $\|\boldsymbol{\xi}_t\|_2 = 1$.

In contrast to the deterministic approach, this method does not necessarily proceed along the *steepest slope* in $(\boldsymbol{x}_t)$, but this property is sometimes a *plus* in the sense that it may *avoid being trapped in* ***local maxima*** or in ***saddlepoints*** of h.

- Consider the objective function is the sample average of log-likelihood function $h(\boldsymbol{\theta}) := \sum_{i=1}^{N} \ell(\boldsymbol{\theta}; \boldsymbol{x}_i)$. The ***stochastic gradient ascent*** is of form

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_t; \boldsymbol{\xi}_t), \quad \alpha_t > 0. \tag{3}$$

where $\boldsymbol{\xi}_t \sim \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is a random sample from data.

2.2 Simulated Annealing

- The ***fundamental idea*** of ***simulated annealing methods*** is that a change of scale, called ***temperature***, allows for faster moves on the surface of the function h to maximize, whose negative is called ***energy***. Therefore, rescaling partially avoids the trapping attraction of local maxima.

- Given a temperature parameter $T > 0$, $(\boldsymbol{X}_t^{(k)})$ is generated from the distribution

$$\pi_k(\boldsymbol{x}) \propto \exp\left(\frac{h(\boldsymbol{x})}{T_k}\right) \tag{4}$$

As T_k decreases toward 0, the values simulated from this distribution become concentrated in a narrower and narrower neighborhood of the *local maxima* of h

- In Simulated Annealing, we run a fixed number of N_k iterations of MCMC (*Metropolis-Hastings* or *Gibbs sampling*) so that $(\boldsymbol{X}_t^{(k)})$ follows the stationary distribution $\pi_k(\boldsymbol{x}) \propto \exp(h(\boldsymbol{x})/T_k)$. Then we decrease temperature T_k to T_{k+1} and restart the MCMC with last update $\boldsymbol{X}_k$. As $T_k \to 0$, the target distribution is approaching to an indicator on its mode, thus the simulated samples would be close to the optimal solution.

- The ***Simulated Annealing algorithm*** can be implemented as

  1. Simulate $\boldsymbol{\xi}$ from an *instrumental distribution* with density $g(\|\boldsymbol{\xi} - \boldsymbol{X}_t\|)$;
  2. Accept $\boldsymbol{X}_{t+1} = \boldsymbol{\xi}$ with probability

$$\alpha(\boldsymbol{X}_t, \boldsymbol{\xi}) = \min\left\{1, \exp\left(\frac{\Delta h_t}{T_t}\right)\right\}$$

  where $\Delta h_t := h(\boldsymbol{\xi}) - h(\boldsymbol{X}_t)$;

3. Otherwise, accept $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$.

4. Update T_t to T_{t+1}.

- Since there is non-zero probability of accepting a new proposal even if it is not local maximal, it allows the algorithm to *escape the attraction* of $\boldsymbol{x}_t$ if $\boldsymbol{x}_t$ is a local maximum of h. The temperature T_k controls the probability of acceptance of non-optimal proposal.
- An important feature of the simulated annealing algorithm is that there exist convergence results in the case of **finite spaces**. [Robert and Casella, 1999]
- It can be shown that the global optimum of $h(\boldsymbol{x})$ can be reached by Simulated Annealing algorithm with probability 1 if the temperature variable T_k **decreases *sufficiently slowly***, i.e., at the order of $\mathcal{O}(\log(L_k) - 1)$, where $L_k = N_1 + \ldots + N_k$, N_k is the number of iterations of MCMC at time t. In practice, however, no one can afford to have such a slow annealing schedule. Most frequently, people use a linear or even exponential temperature decreasing schedule, which can no longer guarantee that the global optimum will be reached [Liu, 2001].

2.3 Simulated Tempering

- ***Simulated Tempering (ST)*** is proposed to let a MCMC scheme move more freely in the state space [Liu, 2001].
- Define a new target distribution, $\pi(\boldsymbol{x}, i) \propto c_i \exp\{-h(\boldsymbol{x})/T_i\}$ on the augmented space $(\boldsymbol{x}, i) \in \mathcal{X} \times I$. Here, the c_i are constants that can be controlled by the user and they should be tuned so that each tempered distribution in the system should have a roughly equal chance to be visited. Ideally, the c_i should be proportional to the reciprocal of the i-th partition function, $Z_i = \int \exp\{-h(\boldsymbol{x})/T_i\}$.
- The ***Simulated Tempering (ST)*** is described as

  1. With the current state $(\boldsymbol{X}_t, i_t) = (\boldsymbol{x}, i)$, we draw $u \sim \mathcal{U}[0, 1]$.
  2. (***Update sample***) If $u \leq \alpha_0$, we let $i_{t+l} = i$ and let $\boldsymbol{X}_{t+1}$ be drawn from a ***MCMC transition*** $K_i(\boldsymbol{x}, \boldsymbol{x}_{t+1})$ that leaves $\pi_i \propto \exp(h(\boldsymbol{x})/T_i)$ invariant.
  3. (***Update temperature***) If $u > \alpha_0$, we let $\boldsymbol{X}_{t+l} = \boldsymbol{x}$ and propose a ***level transition***, $i \to i'$, from a transition function $\alpha(i, i')$ (usually a nearest-neighbor simple random walk with reflecting boundary), and let $i_{t+1} = i'$ with probability
  $$\min\left\{1, \frac{c_{i'}\, \pi_{i'}(\boldsymbol{x})\, \alpha(i', i)}{c_i\, \pi_i(\boldsymbol{x})\, \alpha(i, i')}\right\};$$
  otherwise accept $i_{t+1} = i$.

3 Stochastic Approximations

3.1 The EM Algorithm

- Missing data models are best thought of as models where the likelihood can be expressed as
$$g(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\mathcal{X}} p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{x} \tag{5}$$
We refer to representation (5) as ***demarginalization***.

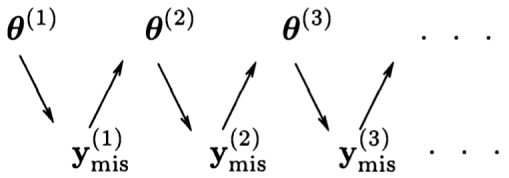

FIGURE 6.2. A graphical illustration of the data augmentation scheme.

Figure 1: Scheme for iteratively fill in missing data and update parameter [Liu, 2001]

- The *incomplete likelihood function* $\ell(\boldsymbol{\theta}|\boldsymbol{x})$ to be optimized can be expressed as the expectation

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) = \mathbb{E}\left[\ell_c(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})\right] = \int_{\mathcal{X}} \ell_c(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x}$$

  We refer to the function $\ell_c(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$ as the *complete-data likelihood*, which corresponds to the observation of the complete data $(\boldsymbol{x}, \boldsymbol{y})$.

- The ***EM (Expectation-Maximization) algorithm*** was originally introduced by Dempster et al. (1977) to overcome the difficulties in maximizing likelihoods by taking advantage of the representation (5) and solving a sequence of easier maximization problems whose limit is the answer to the original problem.

- In particular, we can obtain lower bound of incomplete log-likelihood

$$\begin{aligned}
\log \ell(\boldsymbol{\theta}|\boldsymbol{y}) = \log \int_{\mathcal{X}} p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{x} &= \log \int_{\mathcal{X}} q(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})}{q(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})} d\boldsymbol{x} \\
&\geq \int_{\mathcal{X}} q(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})}{q(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})} d\boldsymbol{x} \\
&= \mathbb{E}_q\left[\ell_c(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})\right] - \mathbb{E}_q\left[q(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})\right]
\end{aligned}$$

  Let $q(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) := \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})}{g(\boldsymbol{y}|\boldsymbol{\theta})}$ be the condition distribution of latent variable given observed data. Common EM notation is to denote the expected log-likelihood by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \boldsymbol{y}) := \mathbb{E}_{\boldsymbol{\theta}_0}\left[\ell_c(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})\right]$$

  where $q(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}_0)$ is the variational distribution.

  Thus the EM algorithm contains two steps:

  1. (E-Step): Compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t, \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{\theta}_t}\left[\ell_c(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})\right]$ with respect to $q(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}_t)$
  2. (M-Step): Optimize q via $\boldsymbol{\theta}_{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t, \boldsymbol{y})$

- It can be shown that EM algorithm increase the incomplete log-likelihood at each iteration:

$$\log \ell(\boldsymbol{\theta}_t|\boldsymbol{y}) \geq \log \ell(\boldsymbol{\theta}_{t-1}|\boldsymbol{y})$$

3.2 Monte Carlo EM

- A difficulty with the implementation of the EM algorithm is that each "E-step" requires the computation of the expected log likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \boldsymbol{y})$. An alternative solution is to

approximate the expectation via Monte Carlo simulation, i.e.

$$\widehat{Q}_m(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \boldsymbol{y}) = \frac{1}{m}\sum_{t=1}^{m} \ell_c(\boldsymbol{\theta}|\boldsymbol{X}_i, \boldsymbol{y}_i) \tag{6}$$

where samples of missing data $\boldsymbol{X}_i \sim q(\boldsymbol{x}|\boldsymbol{y}_i, \boldsymbol{\theta}_0), i = 1, \ldots, m$.

- As $m \to \infty$, $\widehat{Q}_m(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \boldsymbol{y}) \to Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \boldsymbol{y})$.

References

Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.

Christian P Robert and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.