Online Bayesian Moment Matching based SAT Solver Heuristics: Supplemental Material

Haonan Duan [*1,2] **Saeed Nejati** [*1] **George Trimponias** [3] **Pascal Poupart** [1,2] **Vijay Ganesh** [1]

1. BMM for a general SAT instance

In this section, we derive the posterior distribution for an arbitrary clause in a SAT instance and give the pseudocode for BMM on SAT.

Consider an arbitrary clause C, which is a disjunction of a set of literals L. We use m to denote the total number of literals in C, i.e., $|L| = m$. Without loss of generality, we assume that all positive literals appear before negative literals in C, and there are h $(0 \leq h \leq m)$ positive literals.

Based on this, we can express C as follows:

$$C = (\bigvee_{0 \leq i < h} l_i) \vee (\bigvee_{h \leq j < m} \neg l_j).$$

We use the random vector $\mathbf{\Theta}$ to represent the probabilities of each literal being true:

$$\mathbf{\Theta} = \{\theta_k : 0 \leq k < m, \theta_k = P(l_k = T)\}.$$

We assign a product of beta distributions as our prior for $\mathbf{\Theta}$:

$$P(\mathbf{\Theta}) = \prod_{0 \leq k < m} Beta(\theta_k; \alpha_k, \beta_k)$$

The posterior after observing clause C is:

$$\begin{aligned}
P(\mathbf{\Theta}|C) &= \frac{1}{P(C)}(P(\mathbf{\Theta})P(C|\mathbf{\Theta})) \\
&= \frac{1}{P(C)}[\prod_{0 \leq k < m} Beta(\theta_k; \alpha_k, \beta_k)(1 - \prod_{0 \leq i < h}(1 - \theta_i) \prod_{h \leq j < m} \theta_j)] \\
&= \frac{1}{P(C)}[\prod_{0 \leq k < m} Beta(\theta_k; \alpha_k, \beta_k) - \prod_{0 \leq i < h} Beta(\theta_i; \alpha_i, \beta_i) \prod_{0 \leq i < h}(1 - \theta_i) \prod_{h \leq j < m} Beta(\theta_j; \alpha_j, \beta_j) \prod_{h \leq j < m} \theta_j] \\
&= \frac{1}{P(C)}[\prod_{0 \leq k < m} Beta(\theta_k; \alpha_k, \beta_k) - \prod_{0 \leq i < h} Beta(\theta_i; \alpha_i, \beta_i)(1 - \theta_i) \prod_{h \leq j < m} Beta(\theta_j; \alpha_j, \beta_j)\theta_j] \\
&= \frac{1}{P(C)}[\prod_{0 \leq k < m} Beta(\theta_k; \alpha_k, \beta_k) - \prod_{0 \leq i < h} \frac{1}{B(\alpha_i, \beta_i)}\theta_i^{\alpha_i - 1}(1 - \theta_i)^{\beta_i} \prod_{h \leq j < m} \frac{1}{B(\alpha_j, \beta_j)}\theta_j^{\alpha_j}(1 - \theta_j)^{\beta_j - 1}] \\
&= \frac{1}{P(C)}[\prod_{0 \leq k < m} Beta(\theta_k; \alpha_k, \beta_k) - \prod_{0 \leq i < h} \frac{B(\alpha_i, \beta_i + 1)}{B(\alpha_i, \beta_i)} Beta(\theta_i; \alpha_i, \beta_i + 1) \prod_{h \leq j < m} \frac{B(\alpha_j + 1, \beta_j)}{B(\alpha_j, \beta_j)} Beta(\theta_j; \alpha_j + 1, \beta_j)]
\end{aligned}$$

[*]Equal contribution [1]University of Waterloo [2]Vector Institute [3]Huawei Noah's Ark Lab. Correspondence to: Pascal Poupart <ppoupart@uwaterloo.ca>, Vijay Ganesh <vijay.ganesh@uwaterloo.ca>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

$$= \frac{1}{P(C)}[\prod_{0\le k<m} Beta(\theta_k;\alpha_k,\beta_k) - \prod_{0\le i<h}\frac{\beta_i}{\alpha_i+\beta_i} Beta(\theta_i;\alpha_i,\beta_i+1)\prod_{h\le j<m}\frac{\alpha_j}{\alpha_j+\beta_j} Beta(\theta_j;\alpha_j+1,\beta_j)]$$

$$= \frac{1}{P(C)}[\prod_{0\le k<m} Beta(\theta_k;\alpha_k,\beta_k) - \prod_{0\le i<h}\frac{\beta_i}{\alpha_i+\beta_i}\prod_{h\le j<m}\frac{\alpha_j}{\alpha_j+\beta_j}\prod_{0\le i<h} Beta(\theta_i;\alpha_i,\beta_i+1)\prod_{h\le j<m} Beta(\theta_j;\alpha_j+1,\beta_j)],$$

where

$$\begin{aligned}
P(C) &= \int_{(0,1)^m} P(\mathbf{\Theta})P(C|\mathbf{\Theta})d\mathbf{\Theta} \\
&= \int_{(0,1)^m} \prod_{0\le k<m} Beta(\theta_k;\alpha_k,\beta_k) - \prod_{0\le i<h}\frac{\beta_i}{\alpha_i+\beta_i}\prod_{h\le j<m}\frac{\alpha_j}{\alpha_j+\beta_j}\prod_{0\le i<h} Beta(\theta_i;\alpha_i,\beta_i+1)\prod_{h\le j<m} Beta(\theta_j;\alpha_j+1,\beta_j)d\mathbf{\Theta} \\
&= \int_{(0,1)^m} \prod_{0\le k<m} Beta(\theta_k;\alpha_k,\beta_k)d\mathbf{\Theta} \\
&- \prod_{0\le i<h}\frac{\beta_i}{\alpha_i+\beta_i}\prod_{h\le j<m}\frac{\alpha_j}{\alpha_j+\beta_j}\int_{(0,1)^m}\prod_{0\le i<h} Beta(\theta_i;\alpha_i,\beta_i+1)\prod_{h\le j<m} Beta(\theta_j;\alpha_j+1,\beta_j)d\mathbf{\Theta} \\
&= 1 - \prod_{0\le i<h}\frac{\beta_i}{\alpha_i+\beta_i}\prod_{h\le j<m}\frac{\alpha_j}{\alpha_j+\beta_j}.
\end{aligned}$$

We can thus write

$$P(C) = 1-p, \text{ where } p = \prod_{0\le i<h}\frac{\beta_i}{\alpha_i+\beta_i}\prod_{h\le j<m}\frac{\alpha_j}{\alpha_j+\beta_j}.$$

Note that the likelihood $P(C|\mathbf{\Theta})$ can also be calculated as sums of 2^m-1 joint probabilities. We observe that the posterior is a mixture $P(\mathbf{\Theta}|C)$ of products of Beta distributions. The number of mixtures grows exponentially as more clauses are encountered. To address this, we use BMM to approximate the true mixture $P(\mathbf{\Theta}|C)$ by a single product of Beta distributions:

$$\tilde{P}(\tilde{\mathbf{\Theta}}) = \prod_{0\le k<m} Beta(\tilde{\theta}_k;\tilde{\alpha}_k,\tilde{\beta}_k).$$

The parameters $\tilde{\alpha}_k, \tilde{\beta}_k$ for literal l_k are then computed by matching the first and second moments of the marginal distribution of θ_k (we proceed similarly for other literals):

$$\begin{cases} \mathbb{E}_{\tilde{\theta}_k\sim Beta(\tilde{\theta}_k;\tilde{\alpha}_k,\tilde{\beta}_k)}[\tilde{\theta}_k] = \mathbb{E}_{\theta_k\sim P_{\theta_k}(\theta_k|C)}[\theta_k] \\ \mathbb{E}_{\tilde{\theta}_k\sim Beta(\tilde{\theta}_k;\tilde{\alpha}_k,\tilde{\beta}_k)}[\tilde{\theta}_k^2] = \mathbb{E}_{\theta_k\sim P_{\theta_k}(\theta_k|C)}[\theta_k^2] \end{cases} \iff \begin{cases} \frac{\tilde{\alpha}_k}{\tilde{\alpha}_k+\tilde{\beta}_k} = \mathbb{E}_{\theta_k\sim P_{\theta_k}(\theta_k|C)}[\theta_k] \\ \frac{\tilde{\alpha}_k(\tilde{\alpha}_k+1)}{(\tilde{\alpha}_k+\tilde{\beta}_k)(\tilde{\alpha}_k+\tilde{\beta}_k+1)} = \mathbb{E}_{\theta_k\sim P_{\theta_k}(\theta_k|C)}[\theta_k^2] \end{cases}.$$

In the above expression we have used the fact that the first moment (mean) of the beta distribution $Beta(\theta;\alpha,\beta)$ is $\frac{\alpha}{\alpha+\beta}$, while its second moment is $\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$ (Johnson et al., 1995).

If the literal l_k is positive in C, then:

$$\begin{aligned}
P_{\theta_k}(\theta_k|C) &= \int_{(0,1)^{m-1}} P(\mathbf{\Theta}|C)d\theta_0...d\theta_{k-1}d\theta_{k+1}...d\theta_{m-1} \\
&= \frac{1}{1-p}[Beta(\theta_k;\alpha_k,\beta_k) - p\cdot Beta(\theta_k;\alpha_k,\beta_k+1)].
\end{aligned}$$

If the literal l_k is negative in C, then:

$$\begin{aligned}
P_{\theta_k}(\theta_k|C) &= \int_{(0,1)^{m-1}} P(\mathbf{\Theta}|C)d\theta_0...d\theta_{k-1}d\theta_{k+1}...d\theta_{m-1} \\
&= \frac{1}{1-p}[Beta(\theta_k;\alpha_k,\beta_k) - p\cdot Beta(\theta_k;\alpha_k+1,\beta_k)].
\end{aligned}$$

We thus get:

$$\mathbb{E}_{\theta_k\sim P_{\theta_k}(\theta_k|C)}[\theta_k] = \begin{cases} \frac{1}{1-p}(\frac{\alpha_k}{\alpha_k+\beta_k} - p\cdot\frac{\alpha_k}{\alpha_k+\beta_k+1}), \text{ if } l_k \text{ is positive in } C \\ \frac{1}{1-p}(\frac{\alpha_k}{\alpha_k+\beta_k} - p\cdot\frac{\alpha_k+1}{\alpha_k+\beta_k+1}), \text{ if } l_k \text{ is negative in } C. \end{cases}$$

Algorithm 1 BMM for general SAT

Output: An assignment to all literals
initialize prior $Beta(\theta_k; \alpha_k, \beta_k)$ for each literal l_k; (we typically initialize α_k and β_k to 0.1)
for $n = 1$ **to** $MaxEpochs$ **do**
  for each clause C **do**
    $p := 1$;
    for each literal l_k in C **do**
      if l_k is positive in C **then**
        $p := p \cdot \frac{\beta_k}{\alpha_k+\beta_k}$;
      else
        $p := p \cdot \frac{\alpha_k}{\alpha_k+\beta_k}$;
      end if
    end for
    for each literal l_k in C **do**
      if l_k is positive in C **then**
        $NewFirstMoment := \frac{1}{1-p}(\frac{\alpha_k}{\alpha_k+\beta_k} - p \cdot \frac{\alpha_k}{\alpha_k+\beta_k+1})$;
        $NewSecondMoment := \frac{1}{1-p}(\frac{\alpha_k(\alpha_k+1)}{(\alpha_k+\beta_k)(\alpha_k+\beta_k+1)} - p \cdot \frac{\alpha_k(\alpha_k+1)}{(\alpha_k+\beta_k+1)(\alpha_k+\beta_k+2)})$;
      else
        $NewFirstMoment := \frac{1}{1-p}(\frac{\alpha_k}{\alpha_k+\beta_k} - p \cdot \frac{\alpha_k+1}{\alpha_k+\beta_k+1})$;
        $NewSecondMoment := \frac{1}{1-p}(\frac{\alpha_k(\alpha_k+1)}{(\alpha_k+\beta_k)(\alpha_k+\beta_k+1)} - p \cdot \frac{(\alpha_k+1)(\alpha_k+2)}{(\alpha_k+\beta_k+1)(\alpha_k+\beta_k+2)})$;
      end if
      Solve the following system of equations to compute the new α_k, β_k :

$$\frac{\alpha_k}{\alpha_k+\beta_k} = NewFirstMoment;$$
$$\frac{\alpha_k(\alpha_k+1)}{(\alpha_k+\beta_k)(\alpha_k+\beta_k+1)} = NewSecondMoment;$$

    end for
  end for
end for
for each literal l_k **do**
  if $\alpha_k > \beta_k$ **then**
    $l_k := T$;
  else
    $l_k := F$;
  end if
end for

Similarly, we get:

$$\mathbb{E}_{\theta_k \sim P_{\theta_k}(\theta_k|C)}[\theta_k^2] = \begin{cases} \frac{1}{1-p}(\frac{\alpha_k(\alpha_k+1)}{(\alpha_k+\beta_k)(\alpha_k+\beta_k+1)} - p \cdot \frac{\alpha_k(\alpha_k+1)}{(\alpha_k+\beta_k+1)(\alpha_k+\beta_k+2)}), & \text{if } l_k \text{ is positive in } C \\ \frac{1}{1-p}(\frac{\alpha_k(\alpha_k+1)}{(\alpha_k+\beta_k)(\alpha_k+\beta_k+1)} - p \cdot \frac{(\alpha_k+1)(\alpha_k+2)}{(\alpha_k+\beta_k+1)(\alpha_k+\beta_k+2)}), & \text{if } l_k \text{ is negative in } C. \end{cases}$$

Based on the above discussion, the pseudocode of BMM for SAT is given in Algorithm 1. The variable $MaxEpochs$ denotes the number of epochs. During one epoch, we visit each clause exactly once. As explained in the main document, we find empirically that as few as 10 epochs suffice for a good initialization.

Table 1. Number of solved instances (out of 200) and average runtime (in seconds) of MapleCOMSPS and MapleLCMDistChronoBT and their variations on SAT race 2019 benchmark. SAT column shows how many of the solved instances were satisfiable.

	Initialization	Total	SAT	Avg. time	Best config.
MapleCOMSPS	Default	120	89	696.310	default
	Random	119	88	732.489	Activity-Polarity
	Survey Propagation	115	84	813.637	Polarity
	Jeroslow-Wang	123	92	712.904	Activity
	BMM	125	94	841.985	Polarity
MapleLCMDist	Default	120	88	604.368	default
	Random	119	89	685.499	Polarity
	Survey Propagation	115	83	946.500	Polarity
	Jeroslow-Wang	120	88	830.279	Activity-Polarity
	BMM	122	92	665.060	Activity

2. Evaluation over SAT Race 2019 Application Instances

2.1. Experimental Setup

All jobs were run on Intel(R) Xeon(R) Gold 6148 CPU at 2.40GHz. We used the main track of the SAT race 2019, which contains instances encoding problems from real-world applications, like verification, graph problems and combinatorics. The SAT race benchmark is partitioned into "new" and "old" subsets, marking newly submitted instances to the competition and re-used instances from the past competitions. We used the "new" subset of the instances containing 200 instances. Time limit for solving each instance was 5000 seconds (the same as SAT competitions) and memory limit was 8GB.

2.2. Solver Descriptions

The solvers that we used to incorporate BMM were MapleCOMSPS (gold/silver medalist of SAT competition 2016/2017) (Liang et al., 2017) and MapleLCMDistChronoBT (winner of SAT competition 2018) (Ryvchin & Nadel, 2018). We used 10 epochs to compute the posterior in the pre-processing phase and 1 epoch for each learned unary and binary clause. MapleLCMDistChronoBT switches between Distance, VSIDS and LRB branching heuristics. We initialized activity scores of all of these heuristics. Similarly we initialized both VSIDS and LRB in MapleCOMSPS.

2.3. Results

Table 1 shows the number of solved instances out of 200 instances by the two solvers described above, comparing BMM with other methods. Unlike the SAT 2018 benchmark, the best performing configuration was different among the initialization methods, which is listed in Table 1. For MapleCOMSPS BMM-polarity was the best configuration, and for MapleLCMDistChronoBT BMM-activity was the best performing configuration. In both of the solvers, BMM-based initializations are the best version of their respective solvers, beating the default version by 5 instances in MapleCOMSPS and 2 instances in MapleLCMDistChronoBT. It should be noted that BMM-based versions solves 5 more satisfiable instances compared to default MapleCOMSPS and 4 more satisfiable instances compared to default MapleLCMDistChronoBT.

3. Consistency proof of BMM for the naïve Bayes model

For convenience, we repeat (from Section 2) the update equation of BMM for μ_{n+1} in the naïve Bayes model and the theorem asserting the consistency of the BMM update. The equation for τ_{n+1} is similar but more complex; see Eq. (8).

$$\mu_{n+1} = \mu_n + \frac{1}{\tau_n + 1}\Big[(\frac{c_1\alpha_n}{c_1\alpha_n + c_2\beta_n} - \mu_n)(1 - \mathcal{I}_{n+1}) + (\frac{(1-c_1)\alpha_n}{(1-c_1)\alpha_n + (1-c_2)\beta_n} - \mu_n)\mathcal{I}_{n+1}\Big]. \tag{1}$$

Theorem 1 *When performing BMM with the first and second moment in the naïve Bayes model, the update in Eq. (1) for the first moment converges almost surely to the true underlying θ:*

$$\Pr(\lim_{n\to\infty} \mu_n = \theta) = 1 \tag{2}$$

3.1. Proof Overview

We interpret BMM for the naïve Bayes model (from Section 2) as a stochastic approximation (SA) problem and use the convergence theorem of the general stochastic approximation (SA) (Chen & Ryzhov, 2020) to derive the consistency result.

The general SA structure is of the form (Chen & Ryzhov, 2020):

$$x_{n+1} = x_n - \psi_n(Q_n(W_{n+1}, x_n) + \zeta_n(W_{n+1}, x_n, \psi_n)) \tag{3}$$

where $(x_n)_{n=0}^{\infty} \in R^m$, $(\psi_n)_{n=0}^{\infty}$ is the step size (deterministic or stochastic), $(W_n)_{n=0}^{\infty}$ is a sequence of random variables, $(Q_n)_{n=0}^{\infty}$ and $(\zeta_n)_{n=0}^{\infty}$ are two sequences of real measurable functions. The function ζ_n corresponds to the bias.

Based on the above notation, the following 2 terms are defined:

$$\mathcal{F}_n = \mathcal{B}(W_1, ..., W_n, x_1, ..., x_n, \psi_1, ..., \psi_n)$$
$$R_n(x) = \mathbb{E}[Q_n(W_{n+1}, x)|\mathcal{F}_n]$$

where $\mathcal{B}$ denotes the Borel sigma-algebra.

Chen and Ryzhov prove that $x_n \to \theta$ almost surely if the following 4 assumptions are met (Chen & Ryzhov, 2020):

1. *For any n, the system of equations $R_n(x)$ has a unique root θ, which doesn't depend on n.*
2. *For n = 1, 2, ... and any $\epsilon > 0$,*
$$\inf_{\|x-\theta\|_2^2 > \epsilon, n \in \mathbb{N}} (x - \theta)^T R_n(x) > 0.$$
3. *There exists positive constants C_1 and C_2 such that:*
   - $\sup_{n \in \mathbb{N}} \mathbb{E}[\|Q_n(W_{n+1}, x)\|_2^2 | \mathcal{F}_n] \leq C_1(1 + \|x - \theta\|_2^2)$
   - $\sup_{n \in \mathbb{N}} \mathbb{E}[\|\zeta_n(W_{n+1}, x, \psi_n)\|_2^2 | \mathcal{F}_n]/\psi_n^2 \leq C_2(1 + \|x - \theta\|_2^2)$

   for all x.
4. $\sum_{n=0}^{\infty} \psi_n = \infty, \sum_{n=0}^{\infty} \psi_n^2 < \infty$.

The above result is also true if we use an explicit projection operator as follows:

$$x_{n+1} = \Pi_H[x_n - \psi_n(Q_n(W_{n+1}, x_n) + \zeta_n(W_{n+1}, x_n, \psi_n))], \tag{4}$$

where $H = [M_{low}, M_{high}]^m$ a closed interval such that $x_0, \theta \in H$. Π_H is a projection operator that ensures the boundedness of the iterates, a widely-used approach in SA convergence theory (Kushner & Yin, 2013). Hence, provided that the 4 above assumptions are met, update Equations (3) or (4) both have the consistency property $x_n \to \theta$.

In the following sections, we'll rewrite the problem in the form of Equation (3) and verify that the 4 assumptions given above hold.

3.2. SA formulation

In our setting, Z represents the binary hidden variable and $\mathcal{I}$ the binary observable variable. Let θ represent the unknown probability of the hidden variable $P(Z = 0)$, the unknown quantity we wish to infer from $\{\mathcal{I}_1, \mathcal{I}_2, \dots\}$ in an online fashion. On the other hand, the conditional distribution of $\mathcal{I}|Z$ is fully known, as below:

$$P(\mathcal{I} = 0|Z = 0) = c_1, P(\mathcal{I} = 0|Z = 1) = c_2$$
$$P(\mathcal{I} = 1|Z = 0) = d_1 = 1 - c_1, P(\mathcal{I} = 1|Z = 1) = d_2 = 1 - c_2.$$

To avoid degenerate edge cases, we assume that $\theta \in (0, 1)$ and $c_1 \neq c_2$. Furthermore, we consider that $c_1, c_2 \in (0, 1)$, which trivially implies that $d_1, d_2 \in (0, 1)$. We choose a beta distribution $Beta(\theta_0; \alpha_0, \beta_0)$ as the initial prior over θ.

After observing n binary $i.i.d.$ observations $\{\mathcal{I}_1, \ldots, \mathcal{I}_n\}$ and performing BMM, we will have an estimate θ_n for θ which is distributed as a Beta distribution $Beta(\theta_n; \alpha_n, \beta_n)$. The posterior after observing the $(n+1)^{th}$ point $\mathcal{I}_{n+1}$ is:

$$\begin{aligned}
P(\theta_{n+1}|\mathcal{I}_{n+1}=0) &= \frac{P(\theta_n)P(\mathcal{I}_{n+1}=0|\theta_n)}{P(\mathcal{I}_{n+1}=0)} \\
&= \frac{1}{P(\mathcal{I}_{n+1}=0)}\frac{\theta_n^{\alpha_n-1}(1-\theta_n)^{\beta_n-1}}{B(\alpha_n,\beta_n)}(\theta_n c_1 + (1-\theta_n)c_2) \\
&= \frac{1}{P(\mathcal{I}_{n+1}=0)}(c_1\frac{\theta_n^{\alpha_n}(1-\theta_n)^{\beta_n-1}}{B(\alpha_n,\beta_n)} + c_2\frac{\theta_n^{\alpha_n-1}(1-\theta_n)^{\beta_n}}{B(\alpha_n,\beta_n)}) \\
&= a_{n,0}\frac{\theta_n^{\alpha_n}(1-\theta_n)^{\beta_n-1}}{B(\alpha_n+1,\beta_n)} + b_{n,0}\frac{\theta_n^{\alpha_n-1}(1-\theta_n)^{\beta_n}}{B(\alpha_n,\beta_n+1)} \\
P(\theta_{n+1}|\mathcal{I}_{n+1}=1) &= a_{n,1}\frac{\theta_n^{\alpha_n}(1-\theta_n)^{\beta_n-1}}{B(\alpha_n+1,\beta_n)} + b_{n,1}\frac{\theta_n^{\alpha_n-1}(1-\theta_n)^{\beta_n}}{B(\alpha_n,\beta_n+1)},
\end{aligned}$$

where

$$\begin{aligned}
\frac{1}{P(\mathcal{I}_{n+1}=0)} &= \frac{1}{\int_0^1 \frac{1}{B(\alpha_n,\beta_n)}\theta_n^{\alpha_n-1}(1-\theta_n)^{\beta_n-1}(\theta_n c_1 + (1-\theta_n)c_2)d\theta_n} = \frac{\alpha_n+\beta_n}{c_1\alpha_n + c_2\beta_n} \\
a_{n,0} &= \frac{c_1}{P(\mathcal{I}_{n+1}=0)}\frac{B(\alpha_n+1,\beta_n)}{B(\alpha_n,\beta_n)} = \frac{c_1\alpha_n}{c_1\alpha_n+c_2\beta_n} \in (0,1), \text{ since } c_1, c_2 \in (0,1) \wedge \alpha_n, \beta_n > 0 \\
b_{n,0} &= \frac{c_2\beta_n}{c_1\alpha_n+c_2\beta_n} \in (0,1) \\
a_{n,1} &= \frac{d_1\alpha_n}{d_1\alpha_n+d_2\beta_n} \in (0,1) \\
b_{n,1} &= \frac{d_2\beta_n}{d_1\alpha_n+d_2\beta_n} \in (0,1).
\end{aligned} \tag{5}$$

Let θ_n be a function that is distributed according to the Beta distribution $Beta(\theta_n; \alpha_n, \beta_n)$. We use $\mu_n =$ to denote the mean of θ_n, σ_n^2 to denote the variance of $\theta_n = \alpha_n + \beta_n$, τ_n to denote the precision of θ_n, and λ_n to denote $\frac{1}{\tau_n^2\sigma_n^2}$. In particular, the following equations hold (note the first two are standard identities for the Beta distribution):

$$\begin{aligned}
\mu_n &= \frac{\alpha_n}{\alpha_n+\beta_n} \in (0,1) \text{ (since } \alpha_n, \beta_n > 0) \\
\sigma_n &= \frac{\alpha_n\beta_n}{(\alpha_n+\beta_n)^2(\alpha_n+\beta_n+1)} > 0 \\
\tau_n &= \alpha_n + \beta_n > 0 \\
\lambda_n &= \frac{1}{\tau_n^2\sigma_n^2} > 0.
\end{aligned}$$

BMM approximates the mixture posterior $P(\theta_{n+1}|\mathcal{I}_{n+1})$ by a simpler Beta distribution $Beta(\theta_{n+1}; \alpha_{n+1}, \beta_{n+1})$ by matching the first and second moments. The first moment of this beta distribution is $\frac{\alpha_{n+1}}{\alpha_{n+1}+\beta_{n+1}} = \mu_{n+1}$ while its second moment is $\frac{\alpha_{n+1}(\alpha_{n+1}+1)}{(\alpha_{n+1}+\beta_{n+1})(\alpha_{n+1}+\beta_{n+1}+1)}$. By matching the first moments, we get the update equation of μ_{n+1} as follows. The

equation incorporates the two cases that the observed instance I_{n+1} is 0 or 1.

$$\begin{aligned}
\mu_{n+1} &= (a_{n,0}\frac{\alpha_n+1}{\alpha_n+\beta_n+1} + b_{n,0}\frac{\alpha_n}{\alpha_n+\beta_n+1})(1-\mathcal{I}_{n+1}) + (a_{n,1}\frac{\alpha_n+1}{\alpha_n+\beta_n+1} + b_{n,1}\frac{\alpha_n}{\alpha_n+\beta_n+1})\mathcal{I}_{n+1} \\
&= \frac{1}{\tau_n+1}[(a_{n,0}(\alpha_n+1)+b_{n,0}\alpha_n)(1-\mathcal{I}_{n+1}) + (a_{n,1}(\alpha_n+1)+b_{n,1}\alpha_n)\mathcal{I}_{n+1}] \\
&= \frac{1}{\tau_n+1}[(\alpha_n+a_{n,0})(1-\mathcal{I}_{n+1}) + (a_{n,1}+\alpha_n)\mathcal{I}_{n+1}] \\
&= \mu_n + \frac{1}{\tau_n+1}[(\alpha_n+a_{n,0})(1-\mathcal{I}_{n+1}) + (a_{n,1}+\alpha_n)\mathcal{I}_{n+1}] - \mu_n \\
&= \mu_n + \frac{1}{\tau_n+1}[(\alpha_n+a_{n,0})(1-\mathcal{I}_{n+1}) + (a_{n,1}+\alpha_n)\mathcal{I}_{n+1} - \mu_n(\tau_n+1)] \\
&= \mu_n + \frac{1}{\tau_n+1}[a_{n,0}(1-\mathcal{I}_{n+1}) + a_{n,1}\mathcal{I}_{n+1} + \alpha_n\mu_n(\tau_n+1)] \\
&= \mu_n + \frac{1}{\tau_n+1}[(a_{n,0}-\mu_n)(1-\mathcal{I}_{n+1}) + (a_{n,1}-\mu_n)\mathcal{I}_{n+1}] \\
&= \mu_n + \lambda_n[\frac{\sigma_n^2\tau_n^2}{\tau_n+1}(a_{n,0}-\mu_n)(1-\mathcal{I}_{n+1}) + \frac{\sigma_n^2\tau_n^2}{\tau_n+1}(a_{n,1}-\mu_n)\mathcal{I}_{n+1}].
\end{aligned} \tag{6}$$

It is straightforward to show that $0 < \mu_{n+1} < 1$, given $\alpha_n, \beta_n > 0, \forall n$.

By matching the second moments we similarly get the update equation for the second moment:

$$M_{n+1,2} = \frac{(\mu_n\tau_n+1)(\mu_n\tau_n+a_{n,0})}{(\tau_n+1)(\tau_n+2)}(1-\mathcal{I}_{n+1}) + \frac{(\mu_n\tau_n+1)(\mu_n\tau_n+a_{n,0})}{(\tau_n+1)(\tau_n+2)}(\mathcal{I}_{n+1}) \tag{7}$$

But in this setting it is more convenient to use instead the first moment and the precision ($\tau_{n+1} := \alpha_{n+1} + \beta_{n+1}$) (see also (Chen & Ryzhov, 2020)). The update equation for precision is obtained by first solving Eq. (6) and Eq. (7) in terms of α_{n+1} and β_{n+1} using the fact that $\mu_{n+1} = \frac{\alpha_{n+1}}{\alpha_{n+1}+\beta_{n+1}}$ and $M_{n+1,2} = \frac{\alpha_{n+1}(\alpha_{n+1}+1)}{(\alpha_{n+1}+\beta_{n+1})(\alpha_{n+1}+\beta_{n+1}+1)}$:

$$\tau_{n+1} = \tau_n + \frac{A_{n,0}(\tau_n+1) - \tau_n(\mu_n\tau_n+a_{n,0})B_{n,0}}{B_{n,0}(\mu_n\tau_n+a_{n,0})}(1-\mathcal{I}_{n+1}) + \frac{A_{n,1}(\tau_n+1) - \tau_n(\mu_n\tau_n+a_{n,1})B_{n,1}}{B_{n,1}(\mu_n\tau_n+a_{n,1})}(\mathcal{I}_{n+1}), \tag{8}$$

where we have defined:

$$\begin{aligned}
A_{n,0} &= (\mu_n\tau_n+a_{n,0})^2(\tau_n+2) - (\mu_n\tau_n+a_{n,0})(\mu_n\tau_n+1)(\mu_n\tau_n+2a_{n,0}) \\
B_{n,0} &= (\mu_n\tau_n+1)(\mu_n\tau_n+2a_{n,0})(\tau_n+1) - (\tau_n+2)(\mu_n\tau_n+a_{n,0})^2 \\
A_{n,1} &= (\mu_n\tau_n+a_{n,1})^2(\tau_n+2) - (\mu_n\tau_n+a_{n,1})(\mu_n\tau_n+1)(\mu_n\tau_n+2a_{n,1}) \\
B_{n,1} &= (\mu_n\tau_n+1)(\mu_n\tau_n+2a_{n,1})(\tau_n+1) - (\tau_n+2)(\mu_n\tau_n+a_{n,1})^2.
\end{aligned} \tag{9}$$

We finally derive the update equation for the variance below, using the standard formula $\sigma_{n+1}^2 = M_{n+1,2} - \mu_{n+1}^2$ together with Eq. (6) and Eq. (7). We will need this in Section 3.6.

$$\begin{aligned}
\sigma_{n+1}^2 =& \sigma_n^2 + \frac{(\tau_n+1)(\mu_n\tau_n+1)(\mu_n\tau_n+2a_{n,0}) - (\mu_n\tau_n+a_{n,0})^2(\tau_n+2) - \mu_n(1-\mu_n)(\tau_n+1)(\tau_n+2)}{(\tau_n+1)^2(\tau_n+2)}(1-\mathcal{I}_{n+1}) \\
&+ \frac{(\tau_n+1)(\mu_n\tau_n+1)(\mu_n\tau_n+2a_{n,1}) - (\mu_n\tau_n+a_{n,1})^2(\tau_n+2) - \mu_n(1-\mu_n)(\tau_n+1)(\tau_n+2)}{(\tau_n+1)^2(\tau_n+2)}(\mathcal{I}_{n+1}).
\end{aligned} \tag{10}$$

We next define:

$$\begin{aligned}
E_n &:= \frac{\sigma_n^2\tau_n^2}{\tau_n+1} > 0, \\
Q_n(\mathcal{I}_{n+1}, E_n, a_{n,0}, a_{n,1}, \mu_n) &:= -[E_n(a_{n,0}-\mu_n)(1-\mathcal{I}_{n+1}) + E_n(a_{n,1}-\mu_n)\mathcal{I}_{n+1}].
\end{aligned}$$

Then, the update rule in Equation (6) can be rewritten as:

$$\mu_{n+1} = \mu_n - \lambda_n \cdot Q_n(\mathcal{I}_{n+1}, E_n, a_{n,0}, a_{n,1}, \mu_n),$$

which has the general SA structure (3) with no bias term ($\zeta_n := 0$) and λ_n as the step size ($\psi_n := \lambda_n$).

We further define:

$$\mathcal{F}_n := \mathcal{B}(\mathcal{I}_1, ..., \mathcal{I}_n, \mu_1, ..., \mu_n, \lambda_1, ..., \lambda_n)$$
$$R_n(x) := \mathbb{E}[Q_n(\mathcal{I}_{n+1}, E_n, a_{n,0}, a_{n,1}, x)|\mathcal{F}_n]$$

Also, we assume that the sequences $(\alpha_n)_{n=0}^{\infty}$ and $(\beta_n)_{n=0}^{\infty}$ have positive lower bounds, which is standard in many SA convergence proofs and also consistent with empirical evidence. Then we can show that $\mu_n \to \theta$ almost surely by verifying the 4 assumptions proposed by (Chen & Ryzhov, 2020). The proof for our setting shares elements with Proposition EC.1 for the setting described in Section 4.5.1 of (Chen & Ryzhov, 2020).

3.3. Proof of Assumption 1

In our formulation

$$\begin{aligned} R_n(x) =& \mathbb{E}[Q_n(\mathcal{I}_{n+1}, E_n, a_{n,0}, a_{n,1}, x)|\mathcal{F}_n] \\ =& - E_n\left(\frac{c_1 x}{c_1 x + c_2(1-x)} - x\right)(1 - (\theta d_1 + (1-\theta)d_2)) \\ & - E_n\left(\frac{d_1 x}{d_1 x + d_2(1-x)} - x\right)(\theta d_1 + (1-\theta)d_2)) \end{aligned}$$

it is easy to see that $R_n(\theta) = 0$ for any n. We now show that θ is additionally the unique root for $x \in (0, 1)$. Given that $E_n > 0$, this is equivalent to showing:

$$\left(\frac{c_1}{c_1 x + c_2(1-x)} - 1\right)(1 - (\theta d_1 + (1-\theta)d_2)) + \left(\frac{d_1}{d_1 x + d_2(1-x)} - 1\right)(\theta d_1 + (1-\theta)d_2) = 0$$

Indeed, we have:

$$\begin{aligned} & \left(\frac{c_1}{c_1 x + c_2(1-x)} - 1\right)(1 - (\theta d_1 + (1-\theta)d_2)) + \left(\frac{d_1}{d_1 x + d_2(1-x)} - 1\right)(\theta d_1 + (1-\theta)d_2) = 0 \\ \Rightarrow & \frac{c_1 - c_1 x - c_2(1-x)}{c_1 x + c_2(1-x)}(1 - (\theta d_1 + (1-\theta)d_2)) + \frac{d_1 - d_1 x - d_2(1-x)}{d_1 x + d_2(1-x)}(\theta d_1 + (1-\theta)d_2) = 0 \\ \Rightarrow & \frac{(c_1 - c_2)(1-x)}{c_1 x + c_2(1-x)}(1 - (\theta d_1 + (1-\theta)d_2)) + \frac{(d_1 - d_2)(1-x)}{d_1 x + d_2(1-x)}(\theta d_1 + (1-\theta)d_2) = 0 \end{aligned}$$

Given $c_1 - c_2 = -(d_1 - d_2)$ and $x \in (0, 1)$, the last equation can be written as:

$$\begin{aligned} & \frac{1}{c_1 x + c_2(1-x)}(1 - (\theta d_1 + (1-\theta)d_2)) - \frac{1}{d_1 x + d_2(1-x)}(\theta d_1 + (1-\theta)d_2) = 0 \\ & \Rightarrow \frac{1 - (\theta d_1 + (1-\theta)d_2)}{c_1 x + c_2(1-x)} = \frac{\theta d_1 + (1-\theta)d_2}{d_1 x + d_2(1-x)} \\ & \Rightarrow (1 - (\theta d_1 + (1-\theta)d_2))(d_1 x + d_2(1-x)) = (\theta d_1 + (1-\theta)d_2)(c_1 x + c_2(1-x)) \\ & \Rightarrow (1 - (\theta d_1 + (1-\theta)d_2))(d_1 - d_2)x + (1 - (\theta d_1 + (1-\theta)d_2))d_2 \\ & = (\theta d_1 + (1-\theta)d_2)(c_1 - c_2)x + (\theta d_1 + (1-\theta)d_2)c_2 \\ & \Rightarrow (1 - (\theta d_1 + (1-\theta)d_2))d_2 - (\theta d_1 + (1-\theta)d_2)c_2 = (c_1 - c_2)x \\ & \Rightarrow d_2 - (\theta d_1 + (1-\theta)d_2)d_2 - (\theta d_1 + (1-\theta)d_2)c_2 = -(d_1 - d_2)x \\ & \Rightarrow d_2 - (\theta d_1 + (1-\theta)d_2) = (d_2 - d_1)x \\ & \Rightarrow (d_2 - d_1)\theta = (d_2 - d_1)x \\ & \Rightarrow x = \theta. \end{aligned}$$

Note that the last step in the above derivation is valid since we have assumed $c_1 \neq c_2$, or equivalently, $d_1 \neq d_2$.

Technically, note that there is also a root at $x = 0$, which is not in $(0, 1)$. However, it is not hard to show that $x \cdot R_n(x) < 0$ in the neighborhood of $x = 0$ when $\theta \in (0, 1)$, which implies that Assumption 2 is violated at $x = 0$, and the SA algorithm is repelled from $x = 0$, provided that the initial $x_0 \in (0, 1)$ (Borkar, 2008). Similar arguments hold for $x = 1$. Alternatively, we can use the previously mentioned update form (4) with a projection operator Π_H that projects x_{n+1} into a suitable closed interval $[M_{low}, M_{high}] \subset (0, 1)$, where $0 < M_{low}, M_{high} < 1$, so that $x_0, \theta \in H$ (Kushner & Yin, 2013). In that case, the equation $R_n(x) = 0$ has a sole root in the interval H. Finally, given that $\mu_n \in (0, 1) \forall n$, we safely assume everywhere in the proof that $x \in (0, 1)$.

To simplify the proof, from now on we can assume a projection operator Π_H that projects x_{n+1} into a suitable closed interval $[M_{low}, M_{high}] \subset (0, 1)$, as explained above.

3.4. Proof of Assumption 2

To show that Assumption 2 holds, it is sufficient to show that when $x > \theta$, $R_n(x) > 0$ and when $x < \theta$, $R_n(x) < 0$ (for $x \in (0, 1)$).

Given that $E_n > 0$, it suffices to show that the following expression satisfies the property above:

$$
\begin{aligned}
&-\frac{1}{1-x}[(\frac{c_1}{(c_1x + c_2(1-x))} - 1)(1 - (\theta d_1 + (1-\theta)d_2)) + (\frac{d_1}{(d_1x + d_2(1-x))} - 1)(\theta d_1 + (1-\theta)d_2)] \\
&= -\frac{1}{1-x}[\frac{(c_1 - c_2)(1-x)}{(c_1x + c_2(1-x))})(1 - (\theta d_1 + (1-\theta)d_2)) + (\frac{(d_1 - d_2)(1-x)}{(d_1x + d_2(1-x))})(\theta d_1 + (1-\theta)d_2)] \\
&= -[\frac{(c_1 - c_2)}{(c_1x + c_2(1-x))})(1 - (\theta d_1 + (1-\theta)d_2)) + (\frac{(d_1 - d_2)}{(d_1x + d_2(1-x))})(\theta d_1 + (1-\theta)d_2)] \\
&= -(c_1 - c_2)[\frac{(1 - (\theta d_1 + (1-\theta)d_2))}{(c_1x + c_2(1-x))}) - (\frac{(\theta d_1 + (1-\theta)d_2)}{(d_1x + d_2(1-x))})] \\
&= -(c_1 - c_2)[\frac{(1 - (\theta d_1 + (1-\theta)d_2))}{1 - (d_1x + d_2(1-x))}) - (\frac{(\theta d_1 + (1-\theta)d_2)}{(d_1x + d_2(1-x))})] \\
&= -(c_1 - c_2)\frac{(x-\theta)(d_1 - d_2)}{(1 - (d_1x + d_2(1-x)))((d_1x + d_2(1-x)))} \\
&= \frac{(x-\theta)(d_1 - d_2)^2}{(1 - (d_1x + d_2(1-x)))((d_1x + d_2(1-x)))}
\end{aligned}
$$

Given our assumption that $d_1 \neq d_2$, the last expression suggests that for $x \in (0, 1)$ when $x > \theta$, then $R_n(x) > 0$, and when $x < \theta$, then $R_n(x) < 0$.

3.5. Proof of Assumption 3

Since the bias term is identically 0, it is trivial to show the second inequality of Assumption 3. For the first one, we have:

$$Q_n(\mathcal{I}_{n+1}, E_n, a_{n,0}, a_{n,1}, x) = -[E_n(a_{n,0} - x)(1 - \mathcal{I}_{n+1}) + E_n(a_{n,1} - x)\mathcal{I}_{n+1}].$$

We can show that E_n is upper bounded by $\frac{1}{2}$ for any n as follows:

$$E_n = \frac{\sigma_n^2 \tau_n^2}{\tau_n + 1} = \frac{\alpha_n \beta_n}{(\alpha_n + \beta_n + 1)^2} \leq \frac{\alpha_n \beta_n}{2\alpha_n \beta_n} = \frac{1}{2}.$$

Furthermore, given $x, a_{n,0}, a_{n,1} \in (0, 1) \forall n$, it is easy to see that the terms $\sup_{n \in \mathbb{N}} |Q_n(\mathcal{I}_{n+1}, E_n, a_{n,0}, a_{n,1}, x)|$ and $\sup_{n \in \mathbb{N}} \{Q_n^2(\mathcal{I}_{n+1}, E_n, a_{n,0}, a_{n,1}, x)\}$ are upper bounded.

Consequently, there exists a positive constant C_1 such that:

$$\sup_{n \in \mathbb{N}} \mathbb{E}[Q_n^2(\mathcal{I}_{n+1}, E_n, a_{n,0}, a_{n,1}, x) | \mathcal{F}_n] \leq C_1.$$

3.6. Proof of Assumption 4

We only consider the case where the observed term is $\mathcal{I}_{n+1} = 0$. The expression when $\mathcal{I}_{n+1} = 1$ can be tackled in a similar manner.

By using Eq. (8) and (10) to replace τ_{n+1} and σ^2_{n+1}, and doing the tedious calculations, we get the following:

$$\begin{aligned}\frac{1}{\lambda_{n+1}} - \frac{1}{\lambda_n} =& \sigma^2_{n+1}\tau^2_{n+1} - \sigma^2_n\tau^2_n \\ =& S_{n,1} + S_{n,2} + S_{n,3}\end{aligned} \tag{11}$$

where

$$\begin{aligned}S_{n,1} &= \frac{\mu_n(1-\mu_n)}{\tau_n+1}\left(\frac{A_{n,0}(\tau_n+1) - \tau_n(\mu_n\tau_n + a_{n,0})B_{n,0}}{B_{n,0}(\mu_n\tau_n + a_{n,0})}\right)^2 \\ S_{n,2} &= 2\frac{\mu_n(1-\mu_n)}{\tau_n+1}\tau_n\frac{A_{n,0}(\tau_n+1) - \tau_n(\mu_n\tau_n + a_{n,0})B_{n,0}}{B_{n,0}(\mu_n\tau_n + a_{n,0})} \\ S_{n,3} &= \left[\frac{(\tau_n+1)(\mu_n\tau_n+1)(\mu_n\tau_n+2a_{n,0}) - (\mu_n\tau_n+a_{n,0})^2(\tau_n+2) - \mu_n(1-\mu_n)(\tau_n+1)(\tau_n+2)}{(\tau_n+2)}\right] \\ &\quad \cdot \left(\frac{A_{n,0}}{B_{n,0}(\mu_n\tau_n + a_{n,0})}\right)^2.\end{aligned}$$

To prove Assumption 4, it suffices to show that Equation 11 has a positive upper bound and positive lower bound. Indeed, if there exist positive constants $\gamma_*, \gamma^* > 0$ such that for all n

$$\gamma_* \leq \frac{1}{\lambda_{n+1}} - \frac{1}{\lambda_n} \leq \gamma^*,$$

then we must have by (Chen & Ryzhov, 2020) that

$$\frac{1}{\lambda_0} + n\gamma_* \leq \frac{1}{\lambda_n} \leq \frac{1}{\lambda_0} + n\gamma^*.$$

The last inequality, in turn, implies that (Chen & Ryzhov, 2020)

$$\sum_{n=0}^{\infty}\lambda_n = \infty, \sum_{n=0}^{\infty}\lambda_n^2 < \infty,$$

which is what we want to show.

3.6.1. Positive Upper and Lower Bound

If we add the three terms $S_{n,1}, S_{n,2}, S_{n,3}$ in Eq. (11), it is simple to see that the denominator is positive since $\tau_n \in (0,\infty), \mu_n \in (0,1) \wedge a_{n,0} \in (0,1) \forall n$. Furthermore, if τ_n approaches 0, it is simple to show that the denominator is lower bounded by a positive constant assuming μ_n is projected to a closed interval with a projection operator Π_H, as explained in Assumption 1.

Next, we show that the following expression corresponding to the numerator of the sum is always positive:

$$\begin{aligned}&\mu_n(1-\mu_n)(A_{n,0}(\tau_n+1) \\ &- \tau_n(\mu_n\tau_n + a_{n,0})B_{n,0})^2(\tau_n+2) + 2\mu_n(1-\mu_n)\tau_n(A_{n,0}(\tau_n+2) - \tau_n(\mu_n\tau_n+a_{n,0})B_{n,0})(\tau_n+2)(\mu_n\tau_n+a_{n,0})B_{n,0}) \\ &+ [(3\mu_n^2 - (2+2a_{n,0})\mu_n + 2a_{n,0} - a_{n,0}^2)\tau_n + (2\mu_n^2 - 2\mu_n - 2a_{n,0}^2 + 2a_{n,0})]A_{n,0}^2(\tau_n+1).\end{aligned} \tag{12}$$

Equation (12) can be viewed as a function of 3 free variables: $\mu_n, a_{n,0}, \tau_n$. For simplicity, we let $x := \mu_n, y := a_{n,0}, z := \tau_n$. Then it suffices to show that (12) is positive when $x \in (0,1), y \in (0,1), z \in (0,\infty)$. Note also that $x \neq y$, because we have assumed that $c_1 \neq c_2$.

We can factorize (12) as follows:

$$z^2 * (x * z + 1) * (x * z - z - 1) \\ * (x - y)^2 * (x^2 * z^2 + 2 * x * y * z - x * z^2 + y^2 * z - x * z + 2 * y^2 - 2 * y * z - 2 * y) * (x * z + y)^2.$$

We can immediately see that the terms $z^2, (x * z + 1), (x - y)^2, (x * z + y)^2$ are positive. Hence it remains to show that the term $(x * z - z - 1)(x^2 * z^2 + 2 * x * y * z - x * z^2 + y^2 * z - x * z + 2 * y^2 - 2 * y * z - 2 * y)$ is also positive.

Given $x \in (0, 1), z > 0$, we have that

$$xz - z - 1 < z - z - 1 < 0.$$

It thus remains to show that $(x^2 * z^2 + 2 * x * y * z - x * z^2 + y^2 * z - x * z + 2 * y^2 - 2 * y * z - 2 * y)$ is negative. Let's rewrite this expression in terms of z:

$$x^2 * z^2 + 2 * x * y * z - x * z^2 + y^2 * z - x * z + 2 * y^2 - 2 * y * z - 2 * y \\ = (x^2 - x) * z^2 + (y^2 + (2 * x - 2) * y - x) * z + 2 * y^2 - 2 * y.$$

This is a quadratic function of z, where $z \in (0, \infty)$. Because $x, y \in (0, 1)$, it is easy to verify that the coefficients of degree 2 and degree 0 are negative. We can additionally show that the coefficient of degree 1 is negative as follows:

$$\begin{aligned} & y^2 + (2x - 2)y - x \\ & < y + (2x - 2)y - x \\ & = y + 2xy - 2y - x \\ & = 2xy - (x + y) \\ & < 2xy - (x^2 + y^2) \\ & = -(x - y)^2 \\ & < 0. \end{aligned}$$

We have thus established that the sum $S_{n,1} + S_{n,2} + S_{n,3}$ is always positive. Next, we discuss why it is also bounded above and below by positive constants.

In this direction, we first notice that both the numerator and the denominator can be viewed as polynomials of τ_n (or, z), with coefficients that are functions of μ_n (i.e., x) and $a_{n,0}$ (i.e., y). Furthermore, assuming we do BMM with a projection operator Π_H that projects μ_n into a suitable closed interval $[M_{low}, M_{high}] \subset [0, 1]$, where $0 < M_{low}, M_{high} < 1$ and $\theta \in H$, we can see that the numerator can only be 0 if τ_n is 0. Since we showed that Eq. (11) is positive for $\tau \in (0, \infty)$, in order to show a lower and upper bound it suffices to investigate the cases where $\tau_n \to 0$ or $\tau_n \to \infty$.

We first show that, with probability 1, we cannot have that that $\tau_n \to 0$. From Eq. (8) we get:

$$\tau_{n+1} = \frac{A_{n,0}(\tau_n + 1)}{B_{n,0}(\mu_n \tau_n + a_{n,0})}(1 - \mathcal{I}_{n+1}) + \frac{A_{n,1}(\tau_n + 1)}{B_{n,1}(\mu_n \tau_n + a_{n,1})}(\mathcal{I}_{n+1}). \tag{13}$$

Assuming that $\tau_n \approx 0$, we get the following approximation in the limit after we substitute $A_{n,0}, A_{n,1}, B_{n,0}, B_{n,1}$ by (9):

$$\frac{A_{n,0}(\tau_n + 1)}{B_{n,0}(\mu_n \tau_n + a_{n,0})} \to \frac{1}{2}\left(\frac{\mu_n}{a_{n,0}} + \frac{1 - \mu_n}{1 - a_{n,0}}\right) \cdot \tau_n, \tag{14}$$

$$\frac{A_{n,1}(\tau_n + 1)}{B_{n,1}(\mu_n \tau_n + a_{n,1})} \to \frac{1}{2}\left(\frac{\mu_n}{a_{n,1}} + \frac{1 - \mu_n}{1 - a_{n,1}}\right) \cdot \tau_n. \tag{15}$$

Based on (13), (14) and (15), we then get for $\tau_n \approx 0$:

$$\mathbb{E}[\tau_{n+1} | \mathcal{F}_n] \approx \frac{1}{2} \cdot \left((c_1\theta + c_2(1 - \theta))\left(\frac{\mu_n}{a_{n,0}} + \frac{1 - \mu_n}{1 - a_{n,0}}\right) + (d_1\theta + d_2(1 - \theta))\left(\frac{\mu_n}{a_{n,1}} + \frac{1 - \mu_n}{1 - a_{n,1}}\right)\right) \cdot \tau_n. \tag{16}$$

It holds that $d_1\theta + d_2(1 - \theta) = (1 - c_1)\theta + (1 - c_2)(1 - \theta) = 1 - (c_1\theta + c_2(1 - \theta))$. Without loss of generality, let's assume that $c_1 < c_2$. Given $\theta \in (0, 1)$, we then trivially get for the terms $r = c_1\theta + c_2(1 - \theta)$ and $1 - r = d_1\theta + d_2(1 - \theta)$:

$$c_1 < r < c_2 \wedge 1 - c_2 < 1 - r < 1 - c_1. \tag{17}$$

We then have for the term on the right hand side of Eq. (16):

$$\frac{1}{2}\cdot\Big(r(\frac{\mu_n}{a_{n,0}}+\frac{1-\mu_n}{1-a_{n,0}})+(1-r)(\frac{\mu_n}{a_{n,1}}+\frac{1-\mu_n}{1-a_{n,1}})\Big)=\frac{1}{2}\cdot\Big(r\Big(\frac{\frac{1}{1+\nu}}{\frac{1}{1+\frac{c_2}{c_1}\nu}}+\frac{\frac{\nu}{1+\nu}}{\frac{\nu}{\nu+\frac{c_1}{c_2}}}\Big)+(1-r)\Big(\frac{\frac{1}{1+\nu}}{\frac{1}{1+\frac{1-c_2}{1-c_1}\nu}}+\frac{\frac{\nu}{1+\nu}}{\frac{\nu}{\nu+\frac{1-c_1}{1-c_2}}}\Big)\Big)$$

$$=\frac{1}{2}\cdot\Big(r\Big(\frac{1+\frac{c_2}{c_1}\nu}{1+\nu}+\frac{\nu+\frac{c_1}{c_2}}{1+\nu}\Big)+(1-r)\Big(\frac{1+\frac{1-c_2}{1-c_1}\nu}{1+\nu}+\frac{\nu+\frac{1-c_1}{1-c_2}}{1+\nu}\Big)\Big),\ \text{where}\ \nu=\frac{\beta_n}{\alpha_n}\in(0,\infty). \tag{18}$$

We can now show that the term in (18) is greater than 1:

$$\begin{aligned}
&\frac{1}{2}\cdot\Big(r\Big(\frac{1+\frac{c_2}{c_1}\nu}{1+\nu}+\frac{\nu+\frac{c_1}{c_2}}{1+\nu}\Big)+(1-r)\Big(\frac{1+\frac{1-c_2}{1-c_1}\nu}{1+\nu}+\frac{\nu+\frac{1-c_1}{1-c_2}}{1+\nu}\Big)\Big)>1\Leftrightarrow\\
&\quad r\Big(1+\frac{c_2}{c_1}\nu+\nu+\frac{c_1}{c_2}\Big)+(1-r)\Big(1+\frac{1-c_2}{1-c_1}\nu+\nu+\frac{1-c_1}{1-c_2}\Big)>2\cdot(1+\nu)\Leftrightarrow\\
&\quad\quad r\Big(\frac{c_2}{c_1}\nu+\frac{c_1}{c_2}\Big)+(1-r)\Big(\frac{1-c_2}{1-c_1}\nu+\frac{1-c_1}{1-c_2}\Big)>1+\nu\Leftrightarrow\\
&\quad\quad r\frac{c_1}{c_2}+(1-r)\frac{1-c_1}{1-c_2}+\Big(r\frac{c_2}{c_1}+(1-r)\frac{1-c_2}{1-c_1}\Big)\cdot\nu>1+\nu.
\end{aligned}$$

But the last inequality is true, since $r\frac{c_1}{c_2}+(1-r)\frac{1-c_1}{1-c_2}>1$ and $r\frac{c_2}{c_1}+(1-r)\frac{1-c_2}{1-c_1}>1$; this is easy to show given $c_1<r<c_2$ from our original assumption. In fact, both terms can be bounded away from 1, given c_1,c_2,r are distinct (and fixed). As a result of this, Eq. (16) gives for $\tau_n\approx 0$:

$$\mathbb{E}[\tau_{n+1}|\mathcal{F}_n]>\tau_n. \tag{19}$$

The variance will also be finite, because if we assume that $\tau_n\to 0$, then τ_n must have an upper bound. Furthermore, $\tau_n\to 0$ implies that there exists a positive constant K such that $|\tau_{n+1}-\tau_n|\leq K,\forall n$. However, with these assumptions standard martingale theory suggests that, with probability 1, τ_n does not converge to a zero limit (Hall & Heyde, 1980), which is a contradiction. Indeed, by Doob's decomposition theorem, due to Eq. (19) τ can be decomposed into a martingale M and an integrable predictable process A with $A_0=0$ that is almost surely increasing (Hall & Heyde, 1980). Since M converges to 0 almost surely by the martingale central limit theorem and A is strictly increasing almost surely, it is straightforward to see that, with probability 1, τ_n does not converge to a 0 limit.

Finally, we examine the case where $\tau_n\to\infty$. In this direction, we observe that in Eq. (11) if we expand the sum $S_{n,1}+S_{n,2}+S_{n,3}$, both the numerator and the denominator have the same degree. Given the leading coefficients are positive as well as lower and upper bounded since we use a projection operator Π_H, we conclude that the sum must also be positive as well as upper and lower bounded as $\tau_n\to\infty$.

This concludes our proof that $\mu_n\to\theta$ almost surely in the naïve Bayes setting for $\theta\in(0,1)$.

References

Borkar, V. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

Chen, Y. and Ryzhov, I. O. Technical note–consistency analysis of sequential learning under approximate bayesian inference. *Operations Research*, 68(1):295–307, 2020.

Hall, P. and Heyde, C. *Martingale Limit Theory and its Application*. Academic Press, 1980.

Johnson, N., Kotz, S., and Balakrishnan, N. *Continuous univariate distributions*, volume 2. Wiley & Sons, 2nd edition, 1995.

Kushner, H. and Yin, G. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer New York, 2013.

Liang, J. H., Oh, C., Ganesh, V., Czarnecki, K., and Poupart, P. Maple-comsps lrb vsids and maplecomsps chb vsids. *Proc. of SAT Competition*, pp. 20–21, 2017.

Ryvchin, V. and Nadel, A. Maple_lcm_dist_chronobt: Featuring chronological backtracking. *Proc. of SAT Competition*, pp. 29–29, 2018.