Introduction to Machine Learning

Bayesian Learning

Varun Chandola

April 2, 2019

Outline

Contents

1 Generative Models for Discrete Data

- Let $\mathbf{X}$ represents the data with multiple discrete attributes
- Y represent the class

Most probable class

$$P(Y = c|\mathbf{X} = \mathbf{x}, \boldsymbol{\theta}) \propto P(\mathbf{X} = \mathbf{x}|Y = c, \boldsymbol{\theta})P(Y = c, \boldsymbol{\theta})$$

- $P(\mathbf{X} = \mathbf{x}|Y = c, \boldsymbol{\theta}) = p(\mathbf{x}|y = c, \boldsymbol{\theta})$
- $p(\mathbf{x}|y = c, \boldsymbol{\theta})$ - **class conditional density**
- How is the data distributed for each class?
- I give you a set of numbers (training set D) belonging to a concept
- Choose the most likely hypothesis (concept)
- Assume that numbers are between 1 and 100
- Hypothesis Space ($\mathcal{H}$):
  - All powers of 2
  - All powers of 4
  - All even numbers
  - All prime numbers
  - Numbers close to a fixed number (say 12)
  - $\vdots$
- Why choose *powers of 4* concept over *even numbers* concept for $D = \{1, 4, 16, 64\}$?
- Avoid **suspicious coincidences**
- Choose concept with higher *likelihood*
- What is the likelihood of above D to be generated using the *powers of 4* concept?

- Likelihood for *even numbers* concept?

Let h denote the possible concept (or hypothesis). Let $|h|$ or size of h denote the count of numbers that satisfy h. So for numbers between 1 and 100, $|h_{four}|$ (or powers of 4) will be 4 and for *even numbers*, $|h_{even}| = 50$. Likelihood of a data set with N numbers will be given by:

$$p(D|h) = \left[\frac{1}{|h|}\right]^N$$

For the above example, $P(D|h_{four}) = 1/4^4$ and $P(D|h_{even}) = 1/50^4$. So the powers of 4 hypothesis will be selected.

1.1 Likelihood

- Why choose one hypothesis over other?
- Avoid **suspicious coincidences**
- Choose concept with higher *likelihood*

$$p(D|h) = \prod_{x \in D} p(x|h)$$

- *Log Likelihood*

$$\log p(D|h) = \sum_{x \in D} \log p(x|h)$$

1.2 Adding a Prior

- Inside information about the hypotheses
- Some hypotheses are *more likely* apriori
  - May not be the right hypothesis (**prior can be wrong**)

Note that the prior is specified as a probability distribution over all possible hypotheses, and not on the possible outcomes.

1.3 Posterior

- Revised estimates for h after observing evidence (D) and the prior
- *Posterior* $\propto$ *Likelihood* $\times$ *Prior*

$$p(h|D) = \frac{p(D|h)p(h)}{\sum_{h' \in \mathcal{H}} p(D|h')p(h')}$$

For our numbers example, $p(h|D)$ means computing the posterior probability of h to be one of the 10 hypotheses. Note that the individual likelihoods can be analytically computed using the formulation:

$$p(h|D) = \left[\frac{\mathbb{I}(D \in h)}{|h|}\right]^{|D|}$$

The indicator function $\mathbb{I}(D \in h)$ is 1 if every example in D is "covered" by h. Using the likelihood and the prior, one can compute the posterior for the hypotheses in $\mathcal{H}$. Note that we need the summation of all likelihoods in the denominator.

$$\sum_{h' \in \mathcal{H}} p(D|h')p(h') = 0.0772 \times 10^{-3}$$

	h	Prior	Likelihood	Posterior
1	Even	0.300	1.600e-07	1.403e-04
2	Odd	0.075	0.000e+00	0.000e+00
3	Squares	0.075	1.000e-04	2.192e-02
4	Powers of 2	0.100	4.165e-04	1.217e-01
5	Powers of 4	0.075	3.906e-03	8.562e-01
6	Powers of 16	0.075	0.000e+00	0.000e+00
7	Multiples of 5	0.075	0.000e+00	0.000e+00
8	Multiples of 10	0.075	0.000e+00	0.000e+00
9	Numbers within 20 ± 5	0.075	0.000e+00	0.000e+00
10	All Numbers	0.075	1.000e-08	2.192e-06

***Maximum A Priori* Estimate**

$$\hat{h}_{prior} = \arg\max_h p(h)$$

Maximum Likelihood Estimate (MLE)

$$\begin{aligned}\hat{h}_{MLE} &= \arg\max_h p(D|h) = \arg\max_h \log p(D|h)\\ &= \arg\max_h \sum_{x\in D} \log p(x|H)\end{aligned}$$

Maximum a Posteriori **(MAP) Estimate**

$$\hat{h}_{MAP} = \arg\max_h p(D|h)p(h) = \arg\max_h(\log p(D|h) + \log p(h))$$

- $\hat{h}_{prior}$ - Most likely hypothesis based on prior
- $\hat{h}_{MLE}$ - Most likely hypothesis based on evidence
- $\hat{h}_{MAP}$ - Most likely hypothesis based on posterior

$$\hat{h}_{prior} = \arg\max_h \log p(h)$$

$$\hat{h}_{MLE} = \arg\max_h \log p(D|h)$$

$$\hat{h}_{MAP} = \arg\max_h(\log p(D|h) + \log p(h))$$

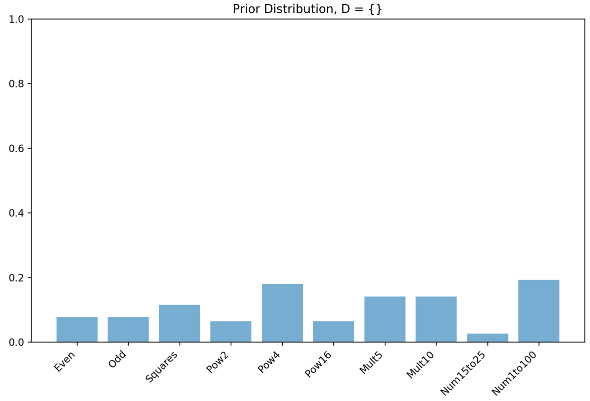
MLE and MAP give the most likely hypothesis before and after considering the prior.

- As data increases, MAP estimate converges towards MLE
  - Why?
- MAP/MLE are **consistent estimators**
  - If concept is in $\mathcal{H}$, MAP/ML estimates will converge
- If $c \notin \mathcal{H}$, MAP/ML estimates converge to h which is closest possible to the truth

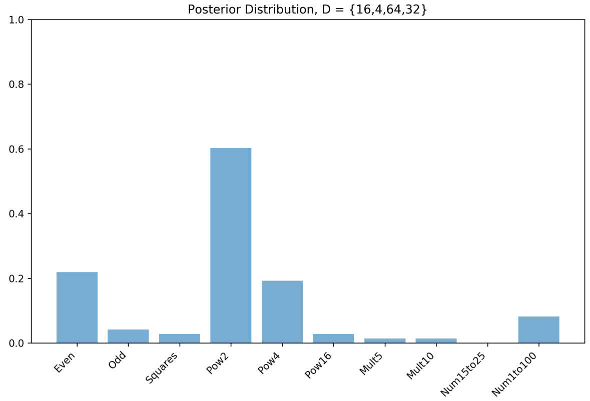
As we have seen in our numbers example, MAP estimate can be written as the sum of log likelihood and log prior for each hypothesis. As data increases, the log likelihood will increase while the log prior will stay constant. Eventually, enough data will overwhelm the prior.

From Prior to Posterior via Likelihood

Prior



Posterior



- Objective: To *revise* the prior distribution over the hypotheses after observing data (evidence).

1.4 Posterior Predictive Distribution

- New input, x^*
- What is the probability that x^* is also generated by the same concept as D?
  - $P(Y = c|X = x^*, D)$?
- **Option 0:** Treat h^{prior} as the true concept

$$P(Y = c|X = x^*, D) = P(X = x^*|c = h^{prior})$$

- **Option 1:** Treat h^{MLE} as the true concept

$$P(Y = c|X = x^*, D) = P(X = x^*|c = h^{MLE})$$

- **Option 2:** Treat h^{MAP} as the true concept

$$P(Y = c|X = x^*, D) = P(X = x^*|c = h^{MAP})$$

- **Option 3:** *Bayesian Averaging*

$$P(Y = c|X = x^*, D) = \sum_h P(X = x^*|c = h)p(h|D)$$

Posterior provides a notion of *belief* about the world. How does one use it? One possible use is to estimate if a new input example belongs to the same concept as the training data, D.

Bayesian averaging assumes that every hypothesis in $\mathcal{H}$ is possible, but with different probabilities. So the output is also a probability distribution.

2 Steps for Learning a Generative Model

- Example: D is a sequence of N binary values (0s and 1s) (coin tosses)
- What is the best distribution that could describe D?

- What is the probability of observing a *head* in future?

Step 1: Choose the form of the model

- Hypothesis Space - All possible distributions
  - Too complicated!!
- Revised hypothesis space - All Bernoulli distributions ($X \sim Ber(\theta), 0 \leq \theta \leq 1$)
  - θ is the hypothesis
  - Still infinite (θ can take infinite possible values)
- Likelihood of D

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

Maximum Likelihood Estimate

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg\max_\theta p(D|\theta) = \arg\max_\theta \theta^{N_1}(1-\theta)^{N_0} \\ &= \frac{N_1}{N_0 + N_1}\end{aligned}$$

To compute MLE we set the derivative of the likelihood with respect to θ to 0.

$$\begin{aligned}\frac{d}{d\theta}\theta_{MLE} &= \frac{d}{d\theta}\theta^{N_1}(1-\theta)^{N_0} \\ &= N_1\theta^{N_1-1}(1-\theta)^{N_0} - N_0\theta^{N_1}(1-\theta)^{N_0-1} \\ &= \theta^{N_1-1}(1-\theta)^{N_0-1}(N_1(1-\theta) - N_0\theta)\end{aligned}$$

Setting above to zero:

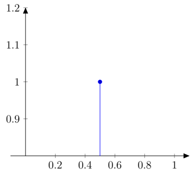
$$\begin{aligned}\theta^{N_1-1}(1-\theta)^{N_0-1}(N_1(1-\theta) - N_0\theta) &= 0 \\ N_1(1-\theta) = N_0\theta \\ \theta = \frac{N_1}{N_0 + N_1}\end{aligned}$$

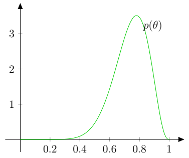
- **We can stop here (MLE approach)**
- Probability of getting a head next:

$$p(x^* = 1|D) = \hat{\theta}_{MLE}$$

2.1 Incorporating Prior

- Prior *encodes* our prior belief on θ
- How to set a Bayesian prior?
  1. A point estimate: $\theta_{prior} = 0.5$
  2. A probability distribution over θ (**a random variable**)
     - Which one?
     - For a bernoulli distribution $0 \leq \theta \leq 1$
     - *Beta* Distribution





2.2 Beta Distribution

- Continuous random variables defined between 0 and 1

$$Beta(\theta|a,b) \triangleq p(\theta|a,b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

- a and b are the (hyper-)parameters for the distribution
- $B(a,b)$ is the **beta function**

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$$

If x is integer

$$\Gamma(x) = (x-1)!$$

- "Control" the shape of the pdf

The *gamma function* is an extension of factorial to real and complex numbers. By varying a and b, one can set any prior on θ, including a uniform prior, a close to point estimate, and a Gaussian prior.

- **We can stop here as well (prior approach)**

$$p(x^* = 1) = \theta_{prior}$$

2.3 Conjugate Priors

- Another reason to choose Beta distribution

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

- Posterior $\propto$ Likelihood × Prior

$$\begin{aligned} p(\theta|D) &\propto \theta^{N_1}(1-\theta)^{N_0}\theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \end{aligned}$$

- **Posterior has same form as the prior**
- Beta distribution is a conjugate prior for Bernoulli/Binomial distribution

Conjugate priors are widely used because they simplify the math and are easy to interpret.

2.4 Estimating Posterior

- Posterior

$$\begin{aligned} p(\theta|D) &\propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \\ &= Beta(\theta|N_1+a, N_0+b) \end{aligned}$$

- We start with a belief that

$$\mathbb{E}[\theta] = \frac{a}{a+b}$$

- After observing N trials in which we observe N_1 heads and N_0 trails, we update our belief as:

$$\mathbb{E}[\theta|D] = \frac{a+N_1}{a+b+N}$$

- We know that posterior over θ is a beta distribution
- MAP estimate

$$\begin{aligned} \hat{\theta}_{MAP} &= \arg\max_{\theta} p(\theta|a+N_1, b+N_0) \\ &= \frac{a+N_1-1}{a+b+N-2} \end{aligned}$$

- What happens if $a = b = 1$?
- **We can stop here as well (MAP approach)**
- Probability of getting a head next:

$$p(x^* = 1|D) = \hat{\theta}_{MAP}$$

Using $a = b = 1$ means using an uninformative prior, which essentially reduces the MAP estimate to MLE estimate.

2.5 Using Predictive Distribution

- All values of θ are possible
- Prediction on an unknown input (x^*) is given by *Bayesian Averaging*

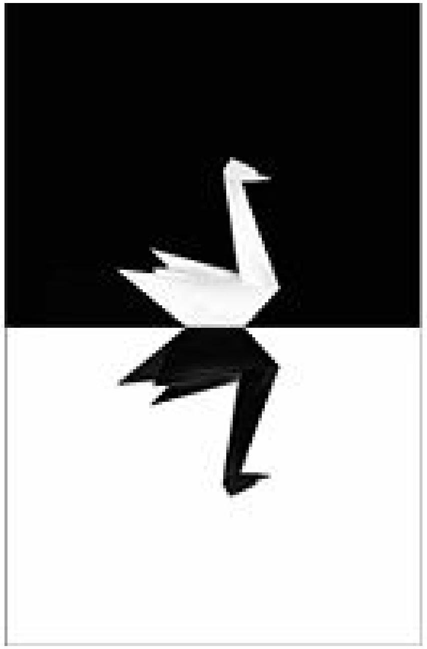
$$\begin{aligned} p(x^* = 1|D) &= \int_0^1 p(x=1|\theta)p(\theta|D)d\theta \\ &= \int_0^1 \theta Beta(\theta|a+N_1, b+N_0) \\ &= \mathbb{E}[\theta|D] \\ &= \frac{a+N_1}{a+b+N} \end{aligned}$$

- This is same as using $\mathbb{E}[\theta|D]$ as a point estimate for θ

2.6 Need for Prior

- Why use a *prior*?
- Consider $D =$ `tails, tails, tails`
- $N_1 = 0, N = 3$
- $\hat{\theta}_{MLE} = 0$
- $p(x^* = 1|D) = 0!!$
  - Never observe a heads
  - The *black swan* paradox
- How does the Bayesian approach help?

$$p(x^* = 1|D) = \frac{a}{a+b+3}$$



The black swan paradox (made famous by an eponymous book by Taleb) essentially states that since one does not observe a phenomenon in the past, he/she incorrectly induces that it can never occur.

2.7 Need for Bayesian Averaging

- MAP is only one part of the posterior
  - θ at which the posterior probability is maximum
  - But is that enough?
  - What about the posterior variance of θ?

$$var[\theta|D] = \frac{(a+N_1)(b+N_0)}{(a+b+N)^2(a+b+N+1)}$$

  - If variance is high then θ_{MAP} is not trustworthy
  - Bayesian averaging helps in this case

3 Learning Gaussian Models

- pdf for MVN with d dimensions:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

3.1 Estimating Parameters

Problem Statement
Given a set of N **independent and identically distributed** (iid) samples, D, learn the parameters $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ of a Gaussian distribution that generated D.

- MLE approach - maximize log-likelihood
- Result

$$\widehat{\boldsymbol{\mu}}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x_i} \triangleq \bar{\mathbf{x}}$$

$$\widehat{\boldsymbol{\Sigma}}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x_i}-\bar{\mathbf{x}})(\mathbf{x_i}-\bar{\mathbf{x}})^\top$$

Proof of the above estimates can be done by maximizing the log-likelihood of the data. The log-likelihood is:

$$\begin{aligned} l(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}}) &= \log p(D|\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}}) \\ &= \frac{N}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x_i}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x_i}-\boldsymbol{\mu}) \\ &= \frac{N}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\sum_{i=1}^{N}\mathbf{y_i}^\top\boldsymbol{\Sigma}^{-1}\mathbf{y_i} \end{aligned}$$

where $\mathbf{y_i} = \mathbf{x_i} - \boldsymbol{\mu}$ (for simplicity).

Maximizing with respect to $\boldsymbol{\mu}$:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}} \mathbf{y_i}^\top \boldsymbol{\Sigma}^{-1}\mathbf{y_i} &= \frac{\partial}{\partial \mathbf{y_i}} \mathbf{y_i}^\top \boldsymbol{\Sigma}^{-1}\mathbf{y_i}\frac{\partial \mathbf{y_i}}{\partial \mu} \\ &= -(\boldsymbol{\Sigma}^{\mathbf{-1}} + \boldsymbol{\Sigma}^{\mathbf{-T}})\mathbf{y_i} \\ &= -(\boldsymbol{\Sigma}^{\mathbf{-1}} + \boldsymbol{\Sigma}^{\mathbf{-T}})(\mathbf{x_i} - \boldsymbol{\mu})\end{aligned}$$

We make of use of a basic result that $\frac{\partial}{\partial \mathbf{x}}\mathbf{x}^\top\mathbf{A}\mathbf{x} = (\mathbf{A}+\mathbf{A}^\top)\mathbf{x}$. Now when we plug the above result to the partial derivative of the log-likelihood with respect to μ:

$$\frac{d}{d\mu} l(\widehat{\mu}, \widehat{\boldsymbol{\Sigma}}) = \frac{1}{2}\sum_{i=1}^{N}(\Sigma^{-1}+\Sigma^{-T})(\mathbf{x_i}-\mu)$$

Setting this to 0 gives us the optimal value of μ:

$$\widehat{\boldsymbol{\mu}}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x_i} \triangleq \bar{\mathbf{x}}$$

Using the optimal value for μ, we can now differentiate the log-likelihood with respect to $\boldsymbol{\Sigma}$. But before that we apply what is known as the "*trace trick*":

$$\mathbf{x}^\top\boldsymbol{\Sigma}^{\mathbf{-1}}\mathbf{x} = tr(\mathbf{x}\mathbf{x}^\top\boldsymbol{\Sigma}^{\mathbf{-1}})$$

and rewrite the log-likelihood in terms of $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{\mathbf{-1}}$:

$$\begin{aligned} l(\boldsymbol{\Lambda}) &= \frac{N}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\sum_{i}^{N} tr[(\mathbf{x_i}-\mu)(\mathbf{x_i}-\mu)^\top\boldsymbol{\Lambda}] \\ &= \frac{N}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}tr[(\sum_{i}^{N}(\mathbf{x_i}-\mu)(\mathbf{x_i}-\mu)^\top)\boldsymbol{\Lambda}] \\ &= \frac{N}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\sum_{i}^{N} tr[\mathbf{S}_\mu\boldsymbol{\Lambda}]\end{aligned}$$

where $\mathbf{S}_\mu \triangleq \sum_i^N (\mathbf{x_i}-\mu)(\mathbf{x_i}-\mu)^\top$ is the *sample scatter matrix*. Making use of the result, $\frac{\partial \log|\mathbf{A}|}{\partial A} = \mathbf{A}^{-\top}$ and $\frac{\partial \mathbf{AB}}{\partial \mathbf{A}} = \mathbf{B}^\top$; differentiating with respect to $\boldsymbol{\Lambda}$:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\Lambda}} l(\boldsymbol{\Lambda}) &= \frac{N}{2}\boldsymbol{\Lambda}^{-\top} - \frac{1}{2}\mathbf{S}_\mu \\ &= \frac{N}{2}\boldsymbol{\Sigma} - \frac{1}{2}\mathbf{S}_\mu\end{aligned}$$

Setting to 0, above equation gives the MLE for covariance matrix as:

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}} &= \frac{1}{N}\mathbf{S}_\mu \\ &= \frac{1}{N}\sum_{i}^{N}(\mathbf{x_i}-\mu)(\mathbf{x_i}-\mu)^\top\end{aligned}$$

3.2 Estimating Posterior

- We need posterior for both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

$$p(\boldsymbol{\mu})$$
$$p(\boldsymbol{\Sigma})$$

- What distribution do we need to sample μ?
  - A Gaussian distribution!

$$p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}_0, \mathbf{V}_0)$$

- What distribution do we need to sample $\boldsymbol{\Sigma}$?
  - An *Inverse-Wishart* distribution.

$$\begin{aligned} p(\boldsymbol{\Sigma}) &= IW(\boldsymbol{\Sigma}|\mathbf{S}, \nu) \\ &= \frac{1}{Z_{IW}}|\boldsymbol{\Sigma}|^{-(\nu+D+1)/2} exp\left(-\frac{1}{2}tr(\mathbf{S}^{-1}\boldsymbol{\Sigma}^{-1})\right)\end{aligned}$$

where,

$$Z_{IW} = |\mathbf{S}|^{-\nu/2}2^{\nu D/2}\Gamma_D(\nu/2)$$

Posterior for $\boldsymbol{\mu}$ - Also a MVN

$$\begin{aligned} p(\boldsymbol{\mu}|D, \boldsymbol{\Sigma}) &= \mathcal{N}(\mathbf{m_N}, \mathbf{V_N}) \\ \mathbf{V}_N^{-1} &= \mathbf{V}_0^{-1} + N\boldsymbol{\Sigma}^{-1} \\ \mathbf{m}_N &= \mathbf{V}_N(\boldsymbol{\Sigma}^{-1}(N\bar{\mathbf{x}}) + \mathbf{V}_0^{-1}\mathbf{m}_0)\end{aligned}$$

Posterior for $\boldsymbol{\Sigma}$ - Also an Inverse Wishart

$$\begin{aligned} p(\boldsymbol{\Sigma}|D, \boldsymbol{\mu}) &= IW(\mathbf{S_N}, \nu_N) \\ \nu_N = \nu_0 + N & \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0 + \mathbf{S}_\mu\end{aligned}$$

References