An Introduction to Stochastic Variational Inference

. . . and some Cauchy talk

WH

May 17, 2023

Systems Neuroscience, TU Dresden

Content

Recap from previous sessions

Recap

Marcov Chain Monte Carlo (MCMC)

- Family of Bayesian methods for sampling from a probability distribution
- Obtain a sequence of random samples which converge to a target distribution
- MCMC methods are relatively slow but asymptotically exact
- Famous methods are *Metropolis–Hastings* and *Hamiltonian Monte Carlo*

Laws of Large Numbers

Weak Law of Large Numbers

Theorem (WLLN)

$$\lim_{n\to\infty} P(|\bar{x}_n - \mu| < \varepsilon) = 1 \qquad \forall \varepsilon \in \mathbb{R}$$

Proof.

$$\lim_{n\to\infty} P(|\bar{x}_n - \mu| \geq \varepsilon) = 0 \qquad \text{equivalent statement} \qquad (1)$$

$$P(|x - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(x)}{\varepsilon^2} \qquad \text{Chebyshev's inequality} \qquad (2)$$

$$P(|\bar{x}_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(\bar{x}_n)}{\varepsilon^2} \qquad \text{substitute } x \text{ by } \bar{x}_n \qquad (3)$$

$$\mathrm{Var}(\bar{x}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} x_i\right) \qquad \mathrm{Var}(ax) = a^2\mathrm{Var}(x) \qquad (4)$$

$$[\dots] \overset{\text{iid}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(x_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma(x_i)^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \qquad \text{equiv transformation} \qquad (5)$$

$$P(|\bar{x}_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(\bar{x}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \qquad 0 \text{ when } n \text{ tends to } \infty \qquad (6)$$

Strong Law of Large Numbers

Theorem (SLLN)

$$P\left(\lim_{n\to\infty} \bar{x}_n = \mu\right) = 1$$

Comments

- **SLLN:** the sample average *converges almost certainly* to expected value
  (sample mean almost surely equals the population mean with infinitely many samples)
- **WLLN:** states that $\bar{x}_n$ *converges in probability* to μ
  (sample mean is increasingly likely to be close to the μ with growing sample size)
- **Almost certainly:** example by picking integer from the interval $[1, 9] \subset \mathbb{R}$
- **Modern proof** of the strong law is more complex than that of the weak law
- There are examples where **WLLN holds but the SLLN does not**

Central Limit Theorem (CLT)

LLNs must not be confused with CLT

Intuitively

For iid random variables, the mean value of these random variables tends towards a normal distribution even if the original variables themselves are not normally distributed.

Lindeberg–Lévy CLT

Let $\{x_1, \ldots, x_n\}$ be a set of i.i.d. random variables with mean μ and finite variance σ^2.
For $n \to \infty$, the random variable $\sqrt{n}\,(\bar{x}_n - \mu)$ *converges in distribution* to $\mathcal{N}(0, \sigma^2)$.

Remarks

- There are various versions of the CLT (classical one is Lindeberg–Lévy CLT)
- In practise, sample sizes of 25 are said to fairly approximate normal distribution

Non convergence of Cauchy distribution

Cauchy distribution

Definition (Cauchy distribution)

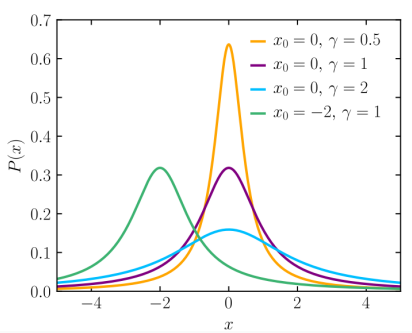
The Cauchy distribution has the following probability density function (pdf)

$$f(x; x_0, \gamma) \coloneqq \frac{1}{\pi}\left[\frac{\gamma}{(x - x_0)^2 + \gamma^2}\right],$$

with x_0 as *location parameter* (symmetric middle), and γ *scale parameter (hwhm)*.

- Mean: undefined
- Median = Mode = x_0
- Variance: undefined
- Entropy: $\log(4\pi\gamma)$
- Moments are generally undefined

Why is this interesting? $\Rightarrow$



Cauchy distribution and non convergence to finite moments

Proof.

1. Sketch proof for first moment (mean) for the Cauchy distribution.
2. For a continuous $\mathbf{pdf}$ $f(x)$, the mean is defined by

$$\mu = \int_{-\infty}^{+\infty} x f(x) dx$$

   - Equivalently, we have $\mu = \int_{-\infty}^{a} [\dots] + \int_{a}^{\infty} [\dots]$ for an arbitrary a.
   - For μ to exist, at least one part should be finite or both infinite in same direction
   - However, taking the antiderivative of the $\mathbf{pdf}$, we obtain the $\mathbf{cdf}$, which is

$$F(x; x_0, \gamma) = \frac{1}{\pi} \arctan\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2}$$

   - In case of Cauchy $\mathbf{pdf}$, both integrals are divergent.
3. Hence, first moment (mean) is undefined and the WLLN cannot be applied. □

Stochastic Variational Inference

Variational Inference

Introductory thoughts

- Alternative to MCMC for **parameter estimation** (approximating intractable integrals) (**vary** estimated distribution to get close to target distribution)
- **Faster** than MCMC, but also less accurate (suitable for large data sets)
- In Bayes: we estimate **posterior** given **joint probability** and not given **evidence**

$$p(\theta \,|\, x) = \frac{p(x \,|\, \theta) \cdot p(\theta)}{p(x)} = \frac{p(x \,|\, \theta) \cdot p(\theta)}{\int_\theta p(x, \theta') d\theta'}$$

- We need two things to do variational inference:
  1. Method to **compare** the **distance** between two probabiliy distributions
  2. Method for **finding extrema** of (loss) functions for optimization
- Such methods are:
  1. **Kullback-Leibler Divergence** (with **ELBO**: evidence lower bound)
  2. **Gradient Descent** / Coordinate Ascent (with **Mean Field Approximation**)

Kullback-Leibler Divergance (DKL)

Comments

- Measure of **"distance"** between two distributions P and Q, denoted by $D_{KL}(Q \parallel P)$
- **Sort of a metric**, but not quite (no symmetry, no triangle inequality)

Definition

Let P and Q be two distributions of a continuous random variable x, and let p and q be their probability densities. The relative entropy D_{KL} is defined as

$$D_{KL}(P \parallel Q) \coloneqq \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

For two Cauchy distributions, we have

$$KL(p_{(x_{0_1},\gamma_1)} || q_{(x_{0_2},\gamma_2)}) \coloneqq \log \frac{(\gamma_1+\gamma_2)^2 + (x_{0_1} - x_{0_2})^2}{4\gamma_1\gamma_2}$$

Evidence Lower Bound (ELBO)

Derivation

$$D_{KL}(q(z) \parallel p(z \mid x)) = \int_z q(z) \log \frac{q(z)}{p(z,x)} \tag{7}$$

$$D_{KL}(q(z) \parallel p(z \mid x)) = \mathbb{E}\left[\log q(z)\right] - \mathbb{E}\left[\log p(z,x)\right] + \log p(x) \tag{8}$$

$$D_{KL}(q(z) \parallel p(z \mid x)) = \mathbb{E}\left[\log q(z)\right] - \mathbb{E}\left[\log p(x \mid z) \cdot p(z)\right] + \log p(x) \tag{9}$$

$$\log p(x) = D_{KL}(q \parallel p) - \mathbb{E}\left[\log q(z) - \log p(z,x)\right] \tag{10}$$

$$\boldsymbol{\log p(x) = D_{KL}(q \parallel p) + \mathcal{L}(q)} \tag{11}$$

Comments

- $\boldsymbol{\mathcal{L}(q)}$ is lower bound for the log-evidence $\log p(x)$ (**ELBO**)
- As $\log p(x)$ is fixed w.r.t q, $\max \mathcal{L}(q) \Rightarrow \min D_{KL}(q \parallel p)$
- Need to know $p(z \mid x)$ only up to **normalizing constant** (numerator, not denominator)
- With suitable choice of q, $\mathcal{L}(q)$ **becomes tractable** (and can hence be maximized)

Gradient Descent

Comments

- General method for **finding extrema of quantitative functions**
- The functions considered to search for extrema are **loss functions** (e.g. sum of the squared residuals in case of linear regression)
- Method can hence be used for **parameter estimation** (esp. when derivation is hard)
- In case of VI, our cost function is the **gradient of the ELBO**

$$\nabla f(x, y) \coloneqq \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right]$$

- Kind of full derivative: **vector points to the direction of steepest descent**

Algorithm (Rough) Gradient Descent Pseudocode

```
1: Pick initial parameter value (guess)
2: Compute loss function; repeat for new (guessed) parameter value
3: Learning rate: define size of steps towards loss function minumum
```