Looking for Convergence Rates for the MAP of the Exponential Family

Anonymous Author
Anonymous Institution

Abstract

We raise the problem of upper bounding the expected sub-optimality of the maximum likelihood estimate, or a conjugate maximum a posteriori for the exponential family. Surprisingly, we found no solution to this problem in the literature – we are not able to tell how many samples we need to fit a gaussian within a few bits of the true distribution. After displaying some properties and special cases of this problem, we show it is a special case of several optimization algorithms, but it falls out of their scopes, thus highlighting range of progress in the analysis of these algorithms.

1 Plan

1. intro to exponential family, and density estimation
2. the thing we want to bound
3. Examples : gaussian mean and gaussian variance (+other examples, just mentioned)
4. Insight : Strongly convex case. (+ self-concordance that is not verified either)
5. insight : bias-variance decomposition
6. Optimization perspective : SBPP or SBG. But no analysis hold, revealing a flaw of all these techniques.
7. Discussion : we believe finding a convergence rate would bring new tools useful to deal with common objects such as barrier losses.

Open questions

Preliminary work. Under review by AISTATS 2022. Do not distribute.

- does a base measure change anything ?
- is there multiple conjugate priors ?
- misspecified case. Do we have a formula ?
- make sure the gaussian entropy is not SC.

2 Introduction and Background

RLP:Goal = disseminate related work to make it look nice. Exponential Families are an elegant and lean way to model a wide variety of data : binary, categorical, natural numbers, positive float, long or short tailed... They are literally the linear model of probabilities. The exponential family for data $X \in \mathcal{X}$ with sufficient statistic T and natural parameter θ is the model

$$p(X|\theta) = \exp(\theta^\top T(X) - A(\theta)) \ , \tag{1}$$

where A is the log-partition function – i.e. the normalization factor

$$A(\theta) = \log \int e^{\theta^\top T(x)} dx \tag{2}$$

where the integral stands for a sum if x has discrete support. RLP:Use base measure instead. Note that an exponential family is entirely specified by its support set $\mathcal{X}$ and its sufficient statistic T. This simple model encompasses both categorical distributions $\mathcal{X} = \{1, \dots, k\}$ with $T(X)$ being the one-hot encoding, and multivariate normal distributions $\mathcal{X} = \mathbb{R}, T(X) = (X, X^2)$.

Duality The logpartition function A verifies the two following identities

$$\nabla A(\theta) = \mathbb{E}_{p(X|\theta)}\left[T(X)\right] =: \mu \tag{3}$$
$$\nabla^2 A(\theta) = \mathrm{Cov}_\theta[T(X)] > 0 \tag{4}$$

where μ is called the mean parameter. If the sufficient statistic T is minimal, then the log-partition function A is strictly convex and its gradient ∇A is a bijection between natural parameters θ and mean parameters μ.

The second identity entails that A is strictly-convex. At this point it is useful to introduce the convex conjugate (aka Fenchel-Legendre transform) of the logpartition function

$$A^*(\mu) = \langle \mu, \theta \rangle - A(\theta) \ . \tag{5}$$

It turns out that A^* matches the common notion of *entropy* in information theory, so we will call it entropy. If A is strictly convex, then its gradient is strictly monotone, so it is a bijection, and its inverse is the gradient of its dual $\nabla A^* \circ \nabla A(\theta) = \theta$ (cf Fig. 1). For a full review of exponential families and their duality, see Wainwright and Jordan (2008, Chapter 3).

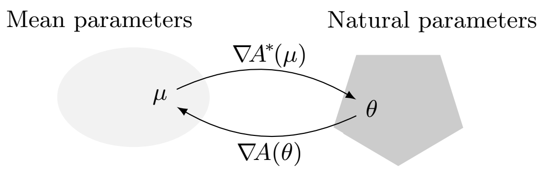

Figure 1: The gradient of the log-partition function and its dual, $(\nabla A, \nabla A^*)$, form a bijection between the natural and mean parameters θ, μ. Figure reproduced from Kunstner et al. (2021).

3 Open Problem

For a well-specified model, the suboptimality on the population log-likelihood is exactly the KL between our current model and the true distribution

$$\mathbb{E}_{X\sim p(.|\theta^*)}\left[-\log p(X|\theta) + \log p(X|\theta^*)\right] = D_{\mathrm{KL}}(p(.|\theta^*); p(.|\theta)) \ . \tag{6}$$

For the exponential family, the KL is also the Bregman divergence induced by the log-partition function (with switched arguments)

$$D_{\mathrm{KL}}(p(.|\theta^*); p(.|\theta)) = \mathcal{B}_A(\theta; \theta^*) \ . \tag{7}$$

There is a general relationship between Bregman divergences and convex conjugates (notice the argument switching)

$$\mathcal{B}_A(\theta; \theta^*) = \mathcal{B}_{A^*}(\mu^*; \mu) \tag{8}$$

so in the end the suboptimality is a divergence, which can either be seen as a KL between distributions, as a divergence between natural parameters, or as a divergence between mean parameters

$$\boxed{D_{\mathrm{KL}}(p(.|\theta^*); p(.|\theta)) = \mathcal{B}_A(\theta; \theta^*) = \mathcal{B}_{A^*}(\mu^*; \mu) \ .} \tag{9}$$

The question is: how does this quantity behave when θ is the maximum-likelihood or the MAP estimate ? Can we get bounds on the following quantities

$$\mathbb{E}_{X_i\sim\theta^*}\left[\mathcal{B}_{A^*}\left(\mathbb{E}[T(X)]; \frac{1}{n}\sum_i T(X_i)\right)\right] \leq ?\ , \tag{10}$$

$$\mathbb{E}_{X_i\sim\theta^*}\left[\mathcal{B}_{A^*}\left(\mathbb{E}[T(X)]; \frac{n_0\mu_0 + \sum_i T(X_i)}{n_0+n}\right)\right] \leq ?\ , \tag{11}$$

where the outer expectation is on the dataset $X_1, \ldots, X_n$?

Remark. What we are looking for is really akin to concentration inequality, expressed with a Bregman divergence instead of a norm. A key difference though, is that the random variable $T(X)$ is connected to the metric A. Indeed expressions (10) or (11) can be infinite for another choice of random variable. For instance, if we plug in $A^*(\mu) = -\log(\mu)$, which defines a divergence on positive numbers, and $T(X) \sim \mathcal{N}(0,1)$ which can be negative.

Remark 2. The expectation of the MLE may be infinite, for instance with $\mathcal{N}(0,\sigma^2)$ and $n \leq 2$. Instead of taking the expectation, we might want to bound this quantity in high probability, without resorting to Markov inequality, but that is a difficult endeavor.

4 Examples

4.1 Gaussian Mean

4.2 Gaussian Variance

The trailing example of this paper is a centered gaussian with unknown variance $\mathcal{N}(0,\sigma^2)$. The density of a centered normal variable is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \ . \tag{12}$$

Defining $T(X) = X^2$ as the sufficient statistic, we get natural parameter $\theta = -\frac{1}{2\sigma^2} < 0$, and mean parameter $\mu = \mathbb{E}[T(X)] = \sigma^2 > 0$. Mean and natural parameters are roughly inverse of each other

$$\theta = -\frac{1}{2\mu} \ . \tag{13}$$

Now we can match the log-likelihood with the exponential family template to get the log-partition function.

$$\log p(x) = -\frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) = x^2\theta - A(\theta) \tag{14}$$

$$\implies A(\theta) = -\frac{1}{2}\log(-\theta) + \frac{1}{2}\log(\pi) \tag{15}$$

We can use the formula $A^*(\mu) = \mu\theta - A(\theta)$ to get the entropy

$$A^*(\mu) = \frac{1}{2}\left(-\log(\mu) + \log\frac{\pi}{2} - 1\right) . \tag{16}$$

We can also take gradient and derivative of $A(\theta)$ to retrieve the mean and covariance of the sufficient statistic X^2

$$\nabla A(\theta) = \frac{-1}{2\theta} = \sigma^2 = \mu = \mathbb{E}[X^2] \tag{17}$$

$$\nabla^2 A(\theta) = \frac{1}{2\theta^2} = 2\sigma^4 = 2\mu^2 = \mathrm{Var}(X^2) \tag{18}$$

which we confirm thank to wikipedia since $\mathbb{E}[X^4] = 3\sigma^4$ and thus $\mathrm{Var}(X^2) = \mathbb{E}[X^4] - \mathbb{E}[X^2]^2 = 3\sigma^4 - \sigma^4 = 2\sigma^4$.

The logpartition function is $A(\theta) = -\log(-\theta)/2 + \mathrm{cst}$, thus the conjugate prior is the exponential family with sufficient statistic $(\theta, \log(-\theta))$, eg a negative Gamma distribution. In particular,

$$p(\theta) \propto \exp(-n_0 A(\theta) + \langle n_0\mu_0, \theta\rangle) \tag{19}$$

$$\propto \exp(\frac{n_0}{2}\log(-\theta) + n_0\mu_0\theta) \tag{20}$$

$$\propto (-\theta)^{1+\frac{n_0}{2}-1} e^{-n_0\mu_0(-\theta)}/Z \tag{21}$$

from which we infer the shape parameter $\alpha = 1 + \frac{n_0}{2}$ and the rate parameter $\beta = n_0\mu_0$, eg $\theta \sim \Gamma(1 + \frac{n_0}{2}, n_0\mu_0)$. After seeing n samples, the posterior is $\Gamma\left(1 + \frac{n_0+n}{2}, n_0\mu_0 + \sum_i T(X_i)\right)$.

Both the entropy and the log-partition are roughly negative logarithm $z \mapsto -\log(z)$. Which yields the same shape of Bregman divergence, as visible below (all three lines are equal)

$$D_{\mathrm{KL}}(\sigma_*^2; \sigma_n^2) = \frac{1}{2}\left(\frac{\sigma_*^2}{\sigma_n^2} - 1 - \log\frac{\sigma_*^2}{\sigma_n^2}\right) \tag{22}$$

$$\mathcal{B}_{A^*}(\mu_*; \mu_n) = \frac{1}{2}\left(\frac{\mu_*}{\mu_n} - 1 - \log\frac{\mu_*}{\mu_n}\right) \tag{23}$$

$$\mathcal{B}_A(\theta_n; \theta_*) = \frac{1}{2}\left(\frac{\theta_n}{\theta_*} - 1 - \log\frac{\theta_n}{\theta_*}\right) . \tag{24}$$

In other words, this divergence measures the discrepancy between the ratio $\frac{\theta_n}{\theta_*} = \frac{\mu_*}{\mu_n}$ and 1 via the function ϕ

$$\phi(z) := \frac{1}{2}(z - 1 - \log(z)) \tag{25}$$

$$\mathcal{B}_A(\theta_n; \theta_*) = \phi(\frac{\theta_n}{\theta_*}) = \phi(\frac{\mu_*}{\mu_n}) \tag{26}$$

as illustrated in Figure 2. We can get a non-transcendental upper bound thanks to the inequality

$$1 - \frac{1}{z} \leq \log(z) \implies \phi(z) \leq \frac{1}{2}(z + \frac{1}{z}) - 1 = \frac{(z-1)^2}{2z} . \tag{27}$$

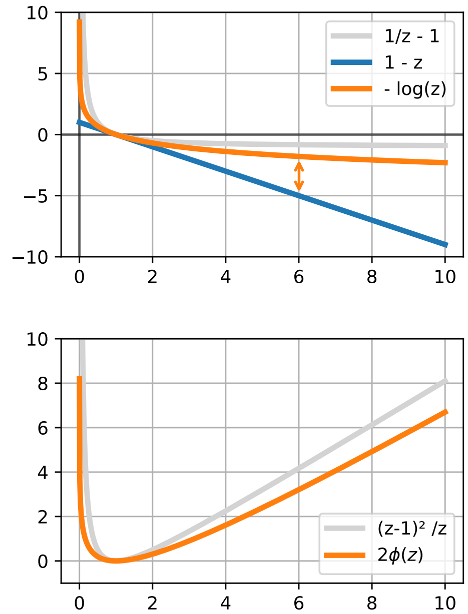

Figure 2: $\phi(z)$ is the Bregman divergence induced by $-\log(z)$. It is a barrier near 0. As a result, it is poorly approximated by quadratics.

References

Kunstner, F., Kumar, R., and Schmidt, M. (2021). Homeomorphic-invariance of EM: Non-asymptotic convergence in KL divergence for exponential families via mirror descent. *AISTATS*.

Wainwright, M. J. and Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

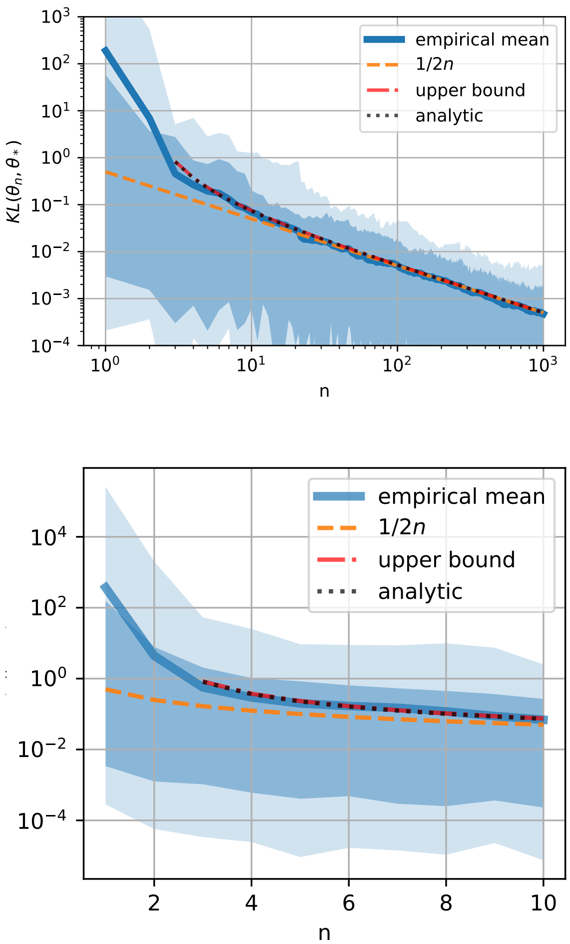

Figure 3: Suboptimality of a Gaussian variance MLE against number of samples n. Bold curve is average over 100 trials, dark shaded area is 90% (dark) confidence interval, light shade is min-max interval. **Left:** as n increases, the suboptimality matches the $1/2N$ asymptote. **Right:** the first few samples significantly deviate from this behavior. In fact, for $n = 1$ and $n = 2$, the expected value is infinite, but we have a closed form solution and a simple upper-bound for $n > 2$.