Convergence rate of MAP estimates for the exponential family

Rémi Le Priol

October 2020

1 Background

It's hard to find general convergence rate on the KL for the maximum likelihood estimates of the exponential family. We want the simplest one. We hope to get a new result by combining tools from statistics and optimization.

The exponential family member with sufficient statistic T and natural parameter θ is the model

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

with log-partition function

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

Recall that 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. The second identity entails that A is strictly-convex. The conjugate prior for this distribution is

$$\pi(\theta) \propto \exp(-n_0 \mathcal{B}_A(\theta||\theta_0)) \tag{5}$$

where n_0 is a number of fictional points observed from a distribution with parameter θ_0. $\mathcal{B}_A(\theta||\theta_0)$ is the Bregman divergence induced by A between θ and θ_0

$$\mathcal{B}_A(\theta||\theta_0) = A(\theta) - A(\theta_0) - \langle \mu_0, \theta - \theta_0 \rangle \tag{6}$$

with $\mu_0 = \nabla A(\theta_0) = \mathbb{E}_{\theta_0}\left[T(X)\right]$ the mean parameter associated to the natural parameter θ_0. The negative log-likelihood of the prior is then

$$-\log \pi(\theta) = n_0(A(\theta) - \theta^\top \mu_0) + \mathrm{cst}$$

Thus the joint NLL of $(x_1, \dots, x_n, \theta)$ is

$$-\log p(X|\theta)\pi(\theta) = (n_0 + n)A(\theta) - \theta^\top \left(n_0\mu_0 + \sum_{i=1}^{n} T(x_i) \right) . \tag{7}$$

Minimizing this expression over θ yields the Maximum A Posteriori estimate

$$\hat{\theta} = \operatorname*{argmin}_\theta - \log p(X|\theta) + n_0 \mathcal{B}_A(\theta||\theta_0) \tag{8}$$

such that the MAP is

$$\nabla A(\hat{\theta}) = \hat{\mu} = \frac{n_0\mu_0 + \sum_{i=1}^n T(X_i)}{n_0 + n} . \tag{9}$$

The MAP estimate is a random quantity.

2 Stochastic Proximal Bregman Point

We see that the MAP estimate minimizes the sum of a stochastic loss $-\log p(X|\theta)$ and a deterministic divergence to an initial point $n_0\mathcal{B}_A(\theta||\theta_0)$. This is a stochastic proximal Bregman step with step-size $\frac{1}{n_0}$. This can also be seen at each step since

$$\begin{aligned} \hat{\theta}_{n+1} &= \operatorname*{argmin}_\theta - \log p(x_{n+1}|\theta) + (n_0 + n)\mathcal{B}_A(\theta||\hat{\theta}_n) && (10)\\ &= \operatorname*{argmin}_\theta f(\theta) + \frac{1}{\gamma_n}\mathcal{B}_A(\theta||\hat{\theta}_n) . && (11) \end{aligned}$$

Hence the MAP estimate can also be seen as the result of a stochastic proximal Bregman point algorithm with step-size $\gamma_n = \frac{1}{n_0+n}$ at step n.

This is similar to the online learning setup, and it may be possible to bound the regret, with approaches similar to Adagrad.

3 Stochastic Mirror Descent

It turns out that the MAP can also be seen as the trajectory of stochastic mirror descent

$$\begin{aligned} \hat{\theta}_{n+1} &= \operatorname*{argmin}_\theta -\langle T(x_{t+1}), \theta\rangle + A(\theta) + (n_0 + n)\mathcal{B}_A(\theta||\theta_n) && (12)\\ &= \operatorname*{argmin}_\theta -\langle T(x_{t+1}), \theta\rangle + A(\theta_n) + \langle \nabla A(\theta_n), \theta - \theta_n\rangle + (n_0 + n + 1)\mathcal{B}_A(\theta||\theta_n) && (13)\\ &= \operatorname*{argmin}_\theta \ell_f(\theta; \theta_n, x_{t+1}) + (n_0 + n + 1)\mathcal{B}_A(\theta||\theta_n) && (14) \end{aligned}$$

where $\ell_f(\theta; \theta_n, x_{t+1})$ is the linearization of f at θ_n evaluated at θ with randomness coming from x_{t+1}. This is the formula for stochastic mirror descent (SMD) applied to f with mirror map A and step-size $\gamma_n = \frac{1}{n_0+n+1}$.

In the classic setting, SMD is studied under strong-convexity assumption on the mirror map A (Bubeck, 2015). In our setting this is not always true – eg gaussians. However a recent and fast-expanding body of work is concerned with a new assumption: relative smoothness and relative strong-convexity. f is μ strongly-convex and L-smooth relative to h if

$$f(x) + \langle \nabla f(x), y - x \rangle + \mu \mathcal{B}_h(y||x) \leq f(y) \tag{15}$$

$$\text{and} \quad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + L\mathcal{B}_h(y||x), \forall x, y \tag{16}$$

$$\iff \mu \mathcal{B}_h(y||x) \leq \mathcal{B}_f(y||x) \leq L\mathcal{B}_h(y||x), \forall x, y \tag{17}$$

$$\iff \mu \nabla^2 h(x) \leq \nabla^2 f(x) \leq L \nabla^2 h(x), \forall x \tag{18}$$

where the last equivalence holds only when f and h are twice differentiable. This sweet generalization of smoothness and strong-convexity transfers the Loewner partial order between matrices to functions, via the Hessian. As such it can be applied to many functions that were out of reach for ℓ^2 norm, by taking the appropriate reference function. For instance $h(x) = -\log(x)$ or $h(x) = x^4$. As early as 2011, Birnbaum et al. (2011) showed $O(\frac{1}{t})$ convergence rate for mirror descent under smoothness assumption relative to the mirror map. More precisely, he proved the simple finite rate

$$f(x_t) - f(x_*) \leq \frac{L\mathcal{B}_h(x_*||x_0)}{t}\,. \tag{19}$$

These notions were rediscovered and expanded by Bauschke et al. (2017) and Lu et al. (2018). If you need to read one, pick Lu et al. (2018) – I found it much much easier and more enjoyable to read. This latter paper also derived a linear convergence rate for mirror descent under relative smoothness and strong-convexity, with the relative condition number $\frac{L}{\mu}$ appearing.

Now our setting is Stochastic Mirror Descent (SMD), meaning at each step we observe a random unbiased estimate gradient. This setting was studied by Hanzely and Richtárik (2018), who proved in the smooth strongly-convex case with tail averaging : with constant step-size, linear convergence down to a variance ball, and with step-size $\gamma_t = n_0 + t$ a rate $\tilde{O}(\frac{1}{t})$. These results match the rates for standard SGD. This is very interesting for us, but the variance hyper-parameter is oddly defined. Let g_t be the random gradient at step t (coming from data point x_t), and θ_{t+1} the next iterate. Then the variance bound σ^2 is a lower bound on the covariance between g_t and θ_{t+1}.

$$\mathrm{Cov}(g_t, \theta_{t+1}) = \mathbb{E}\left[\langle g_t, \theta_{t+1} \rangle\right] - \langle \mathbb{E}[g_t], \mathbb{E}[\theta_{t+1}] \rangle \geq -\gamma_t \sigma^2, \forall t \tag{20}$$

$$= \mathbb{E}\left[\langle g_t - \nabla f(\theta_t), \theta_{t+1} - \theta \rangle\right], (\forall \theta) \tag{21}$$

where γ_t is the step-size and expectations are conditional on θ_t. Still need to check Relative Continuity for Non-Lipschitz Nonsmooth Convex Optimization Using Stochastic (or Deterministic) Mirror Descent, which makes different hypothesis and is well-written.

Also cite "The Information Geometry of Mirror Descent", saying they only give asymptotic results.

4 Straightforward Convergence Rate

In the realizable case, the suboptimality on the population log-likelihood is exactly the KL between our current model and the true distribution

$$\begin{aligned}\mathbb{E}_{x\sim p(.|\theta^*)}\left[-\log p(x|\theta)+\log p(x|\theta^*)\right] &= D_{\mathrm{KL}}(p(.|\theta^*)||p(.|\theta)) && (22)\\ &= \mathcal{B}_A(\theta||\theta^*) && (23)\\ &= \mathcal{B}_{A^*}(\mu^*||\mu) && (24)\end{aligned}$$

where A^* is the entropy, the convex conjugate of the log-partition. The relationship between Bregman divergences and Fenchel conjugacy is well explained in Wainwright's book, and the article geometry of exponential family. The question is: how does this quantity behave when θ is the MAP estimate ? Can we get bounds – in expectation or high-probability ?

If A^* is L-Lipschitz (e.g. A is defined within the ℓ^2-ball of radius L), then

$$\mathcal{B}_{A^*}(\mu^*||\mu) \leq L\|\mu^*-\mu\| + \|\theta\|\|\mu^*-\mu\| \leq 2L\|\mu^*-\mu\| \tag{25}$$

so $\mathcal{B}_{A^*}$ is $2L$-Lipschitz. Since the empirical average converges in expectation to the population mean at a rate of $1/\sqrt{n}$ in ℓ^2 norm, we know that this is the convergence rate of the log-likelihood. RLP:find relevant inequalities.

If A^* is L-smooth (e.g. A is L^{-1}-strongly convex), then

$$\mathcal{B}_{A^*}(\mu^*||\mu) \leq \frac{L}{2}\|\mu^*-\mu\|^2 \tag{26}$$

so $\mathcal{B}_{A^*}$ is upper bounded by a quadratic. In expectation, it should converge at a rate $1/n$.

5 Self-Concordance

A big problem is that A^* is seldom Lipschitz or smooth. For instance the log-partition function of a multivariate normal is

$$A(\eta,\Lambda) = \frac{1}{2}\eta^\top\Lambda^{-1}\eta - \log\det(\Lambda) \tag{27}$$

which is defined on $\eta \in \mathbb{R}^d$ and $\Lambda \in \mathbb{R}^{d\times d}$ symmetric positive definite. It is not strongly convex, so A^* is not smooth.

Another hypothesis that may be more suitable is self-concordance. $f:\mathbb{R}\to\mathbb{R}$ is SC if

$$|f'''(x)| \leq 2f''(x)^{\frac{3}{2}}\ . \tag{28}$$

The exponent $\frac{3}{2}$ is motivated by dimensional analysis and the factor 2 appears to simplify downstream calculus. A multidimensional function $f:\mathbb{R}^n\to\mathbb{R}$ is SC if it's restriction to any line is SC. Negative logarithm $-\log(x)$ and entropy $x\log(x)$ are both self-concordant function. This is good news for us since log-partition function may include SC logarithmic barriers. In particular, gaussians have a logarithmic term. They also have an inverse term which is not self-concordant, but which is generalized self-concordant.

5.1 Fenchel conjugate motivation to self-concordance

In the most regular case, when $f(x)$ is a convex function, continuously differentiable on its domain, then its convex conjugate $f^*(y) = \max_x \langle x, y\rangle - f(x)$ verifies

$$\nabla f \circ \nabla f^* = \mathrm{Id} \tag{29}$$

$$\nabla f^* \circ \nabla f = \mathrm{Id} \tag{30}$$

where Id is the identity function on the relevant domain. In words, the gradients of f and f^* are reciprocal. Deriving this equality yields

$$\nabla^2 f(x) \nabla^2 f^*(x^*) = I_n \tag{31}$$

where x, x^* are conjugate points – e.g. $x^* = \nabla f(x)$ and $x = \nabla f^*(x^*)$. Now, it gets interesting to us when we derive again this equality. Let's tackle the 1D case first

$$f''(x) f^{*\prime\prime}(f'(x)) = 1, \forall x \tag{32}$$

$$\implies f'''(x) f^{*\prime\prime}(f'(x)) + f''(x)^2 f^{*\prime\prime\prime}(f'(x)) = 0 \tag{33}$$

$$\implies \frac{f'''(x)}{f''(x)^{\frac{3}{2}}} + \frac{f^{*\prime\prime\prime}(x^*)}{f^{*\prime\prime}(x^*)^{\frac{3}{2}}} = 0 \tag{34}$$

where to get to the last line we used the first line, and we divided the second line by $f''(x)^{\frac{1}{2}}$. We see that for a pair of conjugate functions, the self-concordance ratio is preserved, modulo the sign. This gives another rational, beyond dimensional analysis, for using this ratio as a regularity assumption for convex analysis.

It is also very helpful for us, since we are looking at pairs A, A^*, and their associated Bregman divergences. If A is SC, then so is $f(\theta) = \mathcal{B}_A(\theta||\theta^*) = A(\theta) - \langle \mu^*, \theta\rangle + \text{cst}$. And A^* is SC as well, thus $h(\mu) = \mathcal{B}_{A^*}(\mu||\mu^*)$ is SC. But there is no reason for $g(\mu) = \mathcal{B}_{A^*}(\mu^*||\mu) = \text{cst} - A^*(\mu) - \langle \nabla A^*(\mu), \mu^* - \mu\rangle$ to be SC.

The multivariate generalization of this formula is a third order tensor equality

$$\nabla^2 f^{-\frac{1}{2}} \nabla^3 f \nabla^2 f^{-1} + \nabla^2 f^{*-\frac{1}{2}} \nabla^3 f^* \nabla^2 f^{*-1} = 0 \tag{35}$$

where we omit multiplication axis and functions take relevant argument x or x^*. Consequently, a multivariate definition of self-concordance might take the form of an inequality on the 3d tensor $\nabla^2 f^{-\frac{1}{2}} \nabla^3 f \nabla^2 f^{-1}$.

5.2 Suboptimality and Newton Decrement

An important property of self-concordant functions (cite Boyd's book, although Nesterov's may be better) is that their suboptimality may be upper bounded by the Newton Decrement

$$D(x)^2 = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \ . \tag{36}$$

In general, subtracting the minimum f^* of f , we have

$$f(x) - f* \leq -D(x) - \log(1 - D(x)) \ . \tag{37}$$

Note that this bound is vacuous for $D(x) \geq 1$. For $y = D(x) \leq 0.68$, we have $-y - \log(1-y) \leq y^2$, so we get the bound

$$f(x) - f^* \leq D(x)^2 = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \ . \tag{38}$$

Our functions of interest is $f(\theta) = \mathcal{B}_A(\theta || \theta^*) = \mathcal{B}_{A^*}(\mu^* || \mu) = g(\mu)$, with minimum $f^* = 0$. If A is self-concordant, then so is f, but not necessarily g. The gradient and Hessian of f are

$$\begin{aligned} f(\theta) &= A(\theta) - A(\theta^*) - \langle \mu^*, \theta - \theta^* \rangle & (39) \\ \nabla f(\theta) &= \mu - \mu^* = \mathbb{E}_{p(X|\theta)}\left[T(X)\right] - \mathbb{E}_{p(X|\theta^*)}\left[T(X)\right] & (40) \\ \nabla^2 f(\theta) &= \Sigma(\theta) = \mathrm{Cov}_{p(X|\theta)}[T(X)] & (41) \end{aligned}$$

so that we get the bound.

$$\mathcal{B}_{A^*}(\mu^* || \mu) \leq D(\theta)^2 = \|\mu^* - \mu(\theta)\|^2_{\Sigma(\theta)^{-1}} \leq 0.46 \tag{42}$$

Finally, if instead we were looking at a different function switching the role of μ and μ^*, $h(\mu) = \mathcal{B}_{A^*}(\mu || \mu^*)$, then we would get

$$\begin{aligned} \nabla h(\mu) &= \theta - \theta^* & (43) \\ \nabla^2 h(\mu) &= \nabla^2 A^*(\mu) = \nabla^2 A(\theta)^{-1} = \mathrm{Cov}_{p(X|\theta)}[T(X)]^{-1} & (44) \\ \implies D(\mu) &= \mathrm{Var}_{p(X|\theta)}[(\theta - \theta^*)^T T(X)] \ . & (45) \end{aligned}$$

This is just a remark. I don't think it can help us to get anywhere.

References

References

Bauschke, H. H., Bolte, J., and Teboulle, M. (2017). A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348.

Birnbaum, B., Devanur, N. R., and Xiao, L. (2011). Distributed algorithms via gradient descent for fisher markets. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 127–136.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

Hanzely, F. and Richtárik, P. (2018). Fastest rates for stochastic mirror descent methods. *arXiv preprint arXiv:1803.07374*.

Lu, H., Freund, R. M., and Nesterov, Y. (2018). Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354.

6 MAP on Graphs

Assume that the variable X factors along some graph G. We write $G(i)$ the parents of X_i in G. Then we model the conditional distribution of X given parameter vector θ factors as

$$p(X|\theta) = \prod_i p(X_i|X_{G(i)};\theta_i) \tag{46}$$

where θ_i is the parameter associated to the mechanism $X_{G(i)} \to X_i$. Embracing the Bayesian viewpoint, the independent mechanism principle is embodied as independence between parameters

$$p(\theta) = \prod_i p(\theta_i)\ . \tag{47}$$

Following these equations, the joint distribution on (θ, X) factors along a larger graph G' which augments G by adding nodes θ_i with arrows pointing to X_i, as illustrated in Figure 1. With such a graph, the Bayesian posterior can be factorized as well

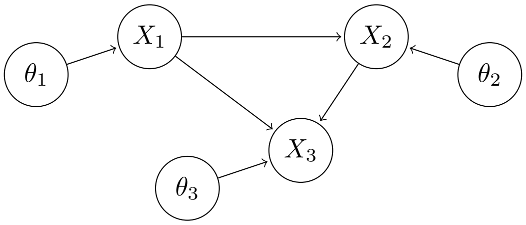

Figure 1: A graph G' factorizing (θ, X). Although the graph restricted on X does not encode any conditional independence, G' does on the joint distribution.

$$\begin{aligned}
p(\theta|X) &\propto p(X|\theta)p(\theta) && (48)\\
&= \prod_i p(X_i|X_{G(i)};\theta_i)p(\theta_i) \qquad (\mathop{\perp\!\!\!\perp}_i \theta_i) && (49)\\
&= \prod_i p(X_i|X_{G(i)};\theta_i)p(\theta_i|X_{G(i)}) \qquad (\theta_i \perp\!\!\!\perp X_{G(i)}) && (50)\\
&= \prod_i p(X_i, \theta_i|X_{G(i)}) && (51)\\
&= \prod_i p(\theta_i|X_i, X_{G(i)})p(X_i|X_{G(i)}) && (52)\\
\implies p(\theta|X) &= \prod_i p(\theta_i|X_i, X_{G(i)})\ . && (53)
\end{aligned}$$

In words, a consequence of the independence mechanism principle is that the posterior distribution of θ_i can be inferred solely from X_i and its parents.

6.1 Equality of directions for 2 categorical variables

In my paper on the analysis of causal speed, I proved the equivalence between sampling a joint distribution $\omega = p(A, B) \in \mathbf{\Delta}_{K\times K}$ on $(A, B) \in \{1, \ldots, K\}^2$ from a Dirichlet with parameter $\gamma \in \mathbb{R}_+^{K\times K}$ and sampling independently the marginal distribution $\mu = p(A) \in \mathbf{\Delta}_K$ and the conditional distributions $\nu_i = p(B|A = i) \in \mathbf{\Delta}_K$ from Dirichlets with respective parameters $\sum_{j=1}^K \gamma_{:,j} = \gamma\,\mathbf{1}$ (matrix vector product) and $\gamma_{i,:}$

$$\underbrace{\mathrm{Dir}_{K^2}((\gamma_{i,j})_{i,j=1}^K)}_{p(\omega)} \equiv \underbrace{\mathrm{Dir}_K\,(\gamma\,\mathbf{1})}_{p(\mu)} \otimes \left(\bigotimes_{i=1}^K \underbrace{\mathrm{Dir}_K((\gamma_{i,j})_j)}_{p(\nu_i)}\right) \tag{54}$$

Seeing data samples $(\mathcal{A}, \mathcal{B}) = (A_i, B_i)_{i=1}^n$ as one-hot encodings in $\mathbb{R}^K \times \mathbb{R}^K$, the posterior reads

$$p(\mu|\mathcal{A}) = \mathrm{Dir}(\gamma\,\mathbf{1} + \sum_i A_i) \tag{55}$$

$$p(\nu_k|\mathcal{A}, \mathcal{B}) = \mathrm{Dir}(\gamma_{k,:} + \sum_i A_{i,k} B_i) \tag{56}$$

$$p(\omega|\mathcal{A}, \mathcal{B}) = \mathrm{Dir}(\gamma + \sum_i A_i B_i^\top) \tag{57}$$

where $A_i B_i^\top$ is the one hot matrix encoding of A, B. These three posteriors are obtained independently of each other following rules of calculus for Dirichlet distributions. Yet they happen to define the same distribution on distributions, as we verify below with the two equalities from equation (54).

$$\left(\gamma + \sum_i A_i B_i^\top\right)_{k,l} = \left(\gamma_{k,:} + \sum_i A_{i,k} B_i\right)_l \tag{58}$$

$$\left(\gamma + \sum_i A_i B_i^\top\right)\mathbf{1} = \gamma\,\mathbf{1} + \sum_i A_i\ . \tag{59}$$

The interpretation of this result is that *taking the posterior with the decomposition $A \to B$ or $B \to A$ give the same result.* As a corollary the MAP is also the same

$$\hat{\omega}^{\mathrm{MAP}} = \frac{\gamma + \sum_i A_i B_i^\top}{\mathbf{1}^\top(\gamma + \sum_i A_i B_i^\top)\,\mathbf{1}} = \frac{\gamma + \sum_i A_i B_i^\top}{n_0 + n} \tag{60}$$

Using Bayesian statistics with this prior, there is no distinction between directions.

Is this bound to happen with a symmetric prior ? Let's give a name to the change of variable $f(\omega) = \mu, \nu$. Remark that $f(\omega^\top) = \mu_\leftarrow, \nu_\leftarrow$, eg in the categorical special case transposing omega and changing variables give the anticausal direction. For sure $p(X|\mu, \nu) = p(X|f(\omega)) = p(X|\omega)$. Using the change of variable formula we get something. But which equality am I looking for exactly ?