Convergence rate of MAP estimates for the exponential family

Rémi Le Priol

October 2020

1 Background

Motivation. We do not know general convergence rates on the KL for 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.

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

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

where A is the log-partition function (aka normalization factor)

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

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

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \ . \tag{3}$$

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

$$\theta = -\frac{1}{2\mu} \ . \tag{4}$$

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

$$\log p(x) = -\frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) = x^2\theta - A(\theta) \tag{5}$$

$$\implies A(\theta) = -\frac{1}{2}\log(-\theta) + \frac{1}{2}\log(\pi) \tag{6}$$

Duality The logpartition function A verifies the two following identities

$$\nabla A(\theta) = \mathbb{E}_{p(X|\theta)}\left[T(X)\right] =: \mu \tag{7}$$

$$\nabla^2 A(\theta) = \mathrm{Cov}_\theta[T(X)] > 0 \tag{8}$$

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

$$A^*(\mu) = \langle \mu, \theta \rangle - A(\theta) \ . \tag{9}$$

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).

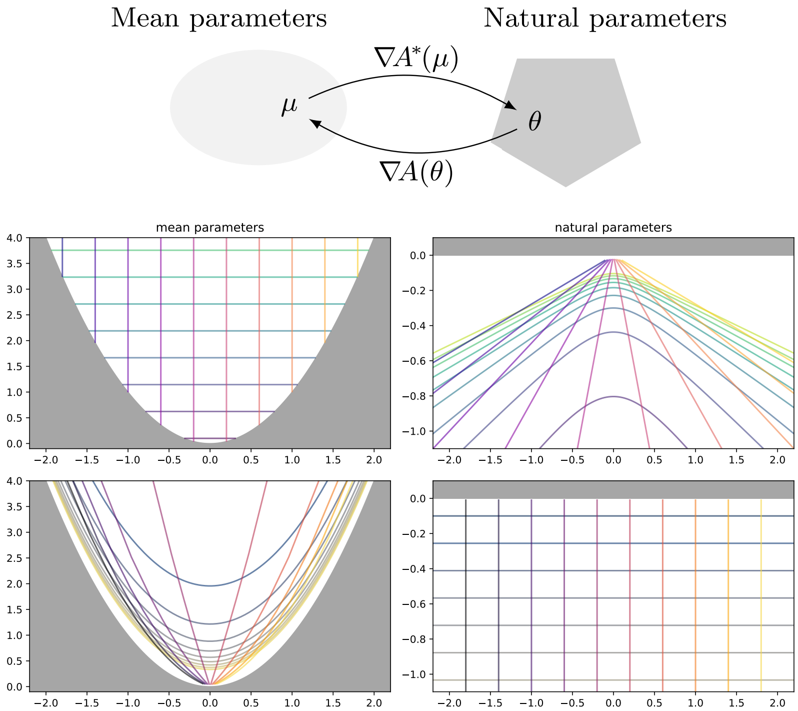

Figure 1: The gradient of the log-partition function and its dual, $(\nabla A, \nabla A^*)$, form a bijection between the natural and mean parameters θ, μ.Top figure reproduced from Kunstner et al. (2021). Bottom figure represents $\mathcal{N}(\mu, \sigma^2)$.

For $\mathcal{N}(0, \sigma^2)$,we can use the formula $A^*(\mu) = \mu\theta - A(\theta)$ to get the entropy

$$A^*(\mu) = \frac{1}{2}\left(-\log(\mu) + \log\frac{\pi}{2} - 1\right) \ . \tag{10}$$

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

$$\nabla A(\theta) = \frac{-1}{2\theta} = \sigma^2 = \mu = \mathbb{E}[X^2] \tag{11}$$

$$\nabla^2 A(\theta) = \frac{1}{2\theta^2} = 2\sigma^4 = 2\mu^2 = \mathrm{Var}(X^2) \tag{12}$$

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

Conjugate Prior One conjugate prior for $p(X|\theta)$ is

$$p(\theta) \propto \exp(-n_0 \mathcal{B}_A(\theta; \theta_0)) \tag{13}$$

where n_0 and θ_0 are (hyper)parameters of the prior, and $\mathcal{B}_A(\theta;\theta_0)$ is the Bregman divergence induced by A between θ and θ_0

$$\mathcal{B}_A(\theta;\theta_0) = A(\theta) - A(\theta_0) - \langle \nabla A(\theta_0), \theta - \theta_0 \rangle \tag{14}$$

with $\nabla A(\theta_0) = \mathbb{E}_{\theta_0}[T(X)] =: \mu_0$ the mean parameter associated to θ_0. Intuitively, n_0 is a number of fictive points observed from a distribution with parameter θ_0. We can re-write this prior as

$$p(\theta) \propto \exp(-n_0 A(\theta) + \langle n_0\mu_0, \theta\rangle) \ , \tag{15}$$

which is the formula for the exponential family with sufficient statistics $(\theta, A(\theta))$ and with natural parameter $(n_0\mu_0, -n_0)$. The posterior given $\mathcal{D} = (X_1, \dots, X_n)$ is then part of the same family, with natural parameters $(n_0\mu_0 + \sum_i T(X_i), -(n_0 + n))$.

In the case of $\mathcal{N}(0, \sigma^2)$, 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,

$$\begin{aligned} p(\theta) &\propto \exp(-n_0 A(\theta) + \langle n_0\mu_0, \theta\rangle) && (16)\\ &\propto \exp(\frac{n_0}{2}\log(-\theta) + n_0\mu_0\theta) && (17)\\ &\propto (-\theta)^{1+\frac{n_0}{2}-1} e^{-n_0\mu_0(-\theta)}/Z && (18) \end{aligned}$$

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 $\Gamma\left(1 + \frac{n_0+n}{2}, n_0\mu_0 + \sum_i T(X_i)\right)$.

Maximum A Posteriori (MAP). The negative log-likelihood of the prior is

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

Thus the joint log-likelihood of $\mathcal{D} = (X_1, \dots, X_n, \theta)$ is

$$-\log p(\mathcal{D}|\theta)p(\theta) = (n_0 + n)A(\theta) - \theta^\top \left(n_0\mu_0 + \sum_{i=1}^n T(X_i) \right) + \mathrm{cst} \ . \tag{19}$$

Minimizing this expression over θ yields the Maximum A Posteriori estimate

$$\hat{\theta} = \operatorname*{argmin}_\theta -\log p(\mathcal{D}|\theta) + n_0\mathcal{B}_A(\theta;\theta_0) \tag{20}$$

such that the MAP is

$$\nabla A(\hat{\theta}_{\mathrm{MAP}}) = \hat{\mu}_{\mathrm{MAP}} = \frac{n_0\mu_0 + \sum_{i=1}^n T(X_i)}{n_0 + n} \ . \tag{21}$$

When $n_0 = 0$ – eg we observed zero samples from the prior – we recover the Maximum Likelihood Estimate (MLE)

$$\hat{\mu}_{\mathrm{MLE}} = \frac{\sum_{i=1}^n T(X_i)}{n} \tag{22}$$

The MLE and MAP estimates are statistics of the dataset $\mathcal{D}$. Given a random dataset, we wish to bound their deviation from the optimum θ^* or μ^*.

2 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

$$\mathbb{E}_{X\sim p(.|\theta^*)}[-\log p(X|\theta) + \log p(X|\theta^*)] = D_{\mathrm{KL}}(p(.|\theta^*); p(.|\theta)) \ . \tag{23}$$

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

$$D_{\mathrm{KL}}(p(.|\theta^*); p(.|\theta)) = \mathcal{B}_A(\theta; \theta^*) \ . \tag{24}$$

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

$$\mathcal{B}_A(\theta; \theta^*) = A(\theta) - \langle \theta, \mu^* \rangle + A^*(\mu^*) = \mathcal{B}_{A^*}(\mu^*; \mu) \tag{25}$$

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

$$\boxed{D_{\mathrm{KL}}(p(.|\theta^*); p(.|\theta)) = \mathcal{B}_A(\theta; \theta^*) = \mathcal{B}_{A^*}(\mu^*; \mu) \ .} \tag{26}$$

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

$$\mathbb{E}_{X_i \sim \theta^*}\left[\mathcal{B}_{A^*}\left(\mathbb{E}[T(X)]; \frac{1}{n}\sum_i T(X_i)\right)\right] \leq ? \ , \tag{27}$$

$$\mathbb{E}_{X_i \sim \theta^*}\left[\mathcal{B}_{A^*}\left(\mathbb{E}[T(X)]; \frac{n_0\mu_0 + \sum_i T(X_i)}{n_0 + n}\right)\right] \leq ? \ , \tag{28}$$

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 (27) or (28) 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.

Below we review all the attempts we made on this problem.

Contents

3 Gaussian Variance Example

For $\mathcal{N}(0, \sigma^2)$, 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)

$$D_{\mathrm{KL}}(\sigma_*^2; \sigma_n^2) = \frac{1}{2}\left(\frac{\sigma_*^2}{\sigma_n^2} - 1 - \log\frac{\sigma_*^2}{\sigma_n^2}\right) \tag{29}$$

$$\mathcal{B}_{A^*}(\mu_*; \mu_n) = \frac{1}{2}\left(\frac{\mu_*}{\mu_n} - 1 - \log\frac{\mu_*}{\mu_n}\right) \tag{30}$$

$$\mathcal{B}_A(\theta_n; \theta_*) = \frac{1}{2}\left(\frac{\theta_n}{\theta_*} - 1 - \log\frac{\theta_n}{\theta_*}\right) . \tag{31}$$

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

$$\phi(z) := \frac{1}{2}(z - 1 - \log(z)) \tag{32}$$

$$\mathcal{B}_A(\theta_n; \theta_*) = \phi(\frac{\theta_n}{\theta_*}) = \phi(\frac{\mu_*}{\mu_n}) \tag{33}$$

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

$$1 - \frac{1}{z} \leq \log(z) \implies \phi(z) \leq \frac{1}{2}(z + \frac{1}{z}) - 1 = \frac{(z-1)^2}{2z} . \tag{34}$$

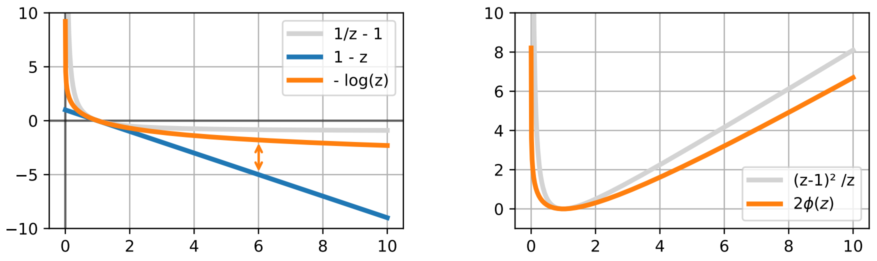

Figure 2: $\phi(z)$ is the Bregman divergence induced by $-\log(z)$. It is a barrier near 0. As a result, it is poorly approximated by quadratics.

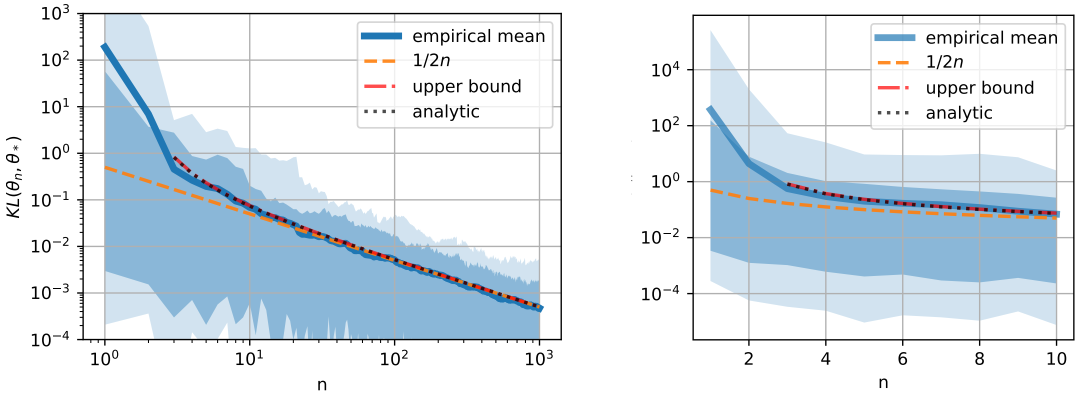

Figure 3: Suboptimality of a Gaussian variance MLE against number of samples n. Bold curve is average over 100 trials, dark shaded area is 90% (dark) confidence interval, light shade is min-max interval. **Left:** as n increases, the suboptimality matches the $1/2N$ asymptote. **Right:** the first few 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 $n > 2$.

Theorem 3.1 (MLE Tight Bound)**.** *The MLE of $\mathcal{N}(0,\mu_*)$ is $\hat\mu_n^{MLE} = \frac{1}{n}\sum_i X_i^2$. Its expected suboptimality is infinite when $n \leq 2$, and otherwise upper-bounded as*

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*;\hat\mu_n^{MLE})\right] \leq \frac{1}{2n} + \frac{2}{n(n-2)}\ . \tag{35}$$

This upper bound is asymptotically tight. We illustrate its behavior against empirical data in Figure 3. There is also a closed form for the multivariate generalization, thanks to the inverse Wishart distribution and the expectation of the log-determinant of a Wishart. The expected value is infinite whenever $n \leq d+1$ where d is the dimension, and we report a similar upper bound in the appendix when $n > d+1$.

As for the MAP, we did not manage to get an asymptotically tight upper bound. Nevertheless, using inequality (34), we did get an interesting upper bound. To start, let us recall that the MAP of $\mathcal{N}(0,\mu_*)$ is $\hat\mu_n = \frac{n_0\mu_0+\sum_i X_i^2}{n_0+n}$. Its expectation is simply $\mu_n = \frac{n_0\mu_0+n\mu^*}{n_0+n}$. We now present a lemma on its expected inverse.

Lemma 3.2 (Expected MAP Natural Parameter)**.** *Let's define $a = n_0\frac{\mu_0}{\mu^*}$. For any $n \geq 1$, the expectation of the natural parameter of the MAP of $\mathcal{N}(0,\mu_*)$ is bounded as*

$$\frac{\mu^*}{\mu_n} \leq \mathbb{E}\left[\frac{\mu^*}{\hat\mu_n}\right] = \mathbb{E}\left[\frac{\hat\theta_n}{\theta^*}\right] \leq \frac{n_0+n}{a+(n-2)_+} \tag{36}$$

where $(x)_+ = \max(0,x)$.

Note that the lower bound is a simple consequence of the convexity of the inverse function. Using this lemma, along with the log upper bound on the logarithm (34), we get a convergence rate for the MAP.

Theorem 3.3 (MAP Bound)**.** *Let us define*

$$b = \frac{(1+\frac{1}{n_0}-\frac{\mu_0}{\mu^*})^2}{2(\frac{\mu_0}{\mu^*}+\frac{(n-2)_+}{n_0})(1+\frac{n}{n_0})}\ . \tag{37}$$

The expected suboptimality of the MAP of $\mathcal{N}(0,\mu^*)$ with prior hyper-parameters (n_0,μ_0) is

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*;\hat\mu_n^{MAP})\right] \leq \begin{cases} \frac{1}{2(n_0+1)} + b \text{ if } n = 1, \\ \frac{1}{n_0\frac{\mu_0}{\mu^*}+n-2} + b \text{ if } n \geq 2 \end{cases} \tag{38}$$

This inequality highlights a clear variance-bias decomposition. In particular, there is no bias term when $\frac{\mu_0}{\mu^*} = 1 + \frac{1}{n_0}$, which happens when the prior is slightly larger than the ground truth. For instance, when $n_0 = 1$, it encourages us to set $\mu_0 = 2\mu^*$. Remark that the variance term is not asymptotically tight as the log-inequality we used (34) is not quadratically tight around 1. We are basically losing a factor 2 compared to $1/2n$.

Using the same bound on the log (34), we get a loose bound on the MLE, which is useful for comparison purposes. We relate these loose convergence rates of MAP and MLE in Figure 4.

Corollary 3.4 (MLE Loose Bound)**.** *When $n \geq 2$, the MLE expected loss is upper bounded as*

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*;\hat\mu_n^{MLE})\right] \leq \frac{1}{n-2}\ . \tag{39}$$

4 Euclidean Behaviour

Often we can relate this Bregman suboptimality with the ℓ^2 distance in mean parameter space. TODO Add base measure to add here a discussion of the euclidean case – the mean of a gaussian

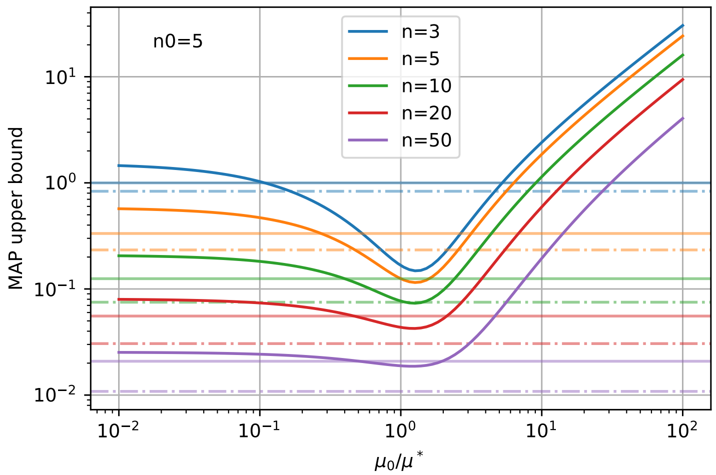

Figure 4: Convergence rate of MAP (38) against the ratio between initialization and optimum $\frac{\mu_0}{\mu^*}$. We also report the tight convergence rate of MLE (35) (dash-dotted horizontal lines), and the loose upper bound (39) (horizontal lines). We report these curves for various values of n, and $n_0 = 5$. We observe that the rate of MAP is significantly better (lower) than the rate of MLE, only when we guessed μ_0 right, within an order of magnitude from μ^*, and for a small number of sample ($n < 10$). This observation hints towards the optimality of MLE for large n. It makes me feel very interested in convergence rates of MAP in the overparametrized regime $n < \dim(T(X))$.

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

$$\mathcal{B}_{A^*}(\mu^*;\mu) \leq L\|\mu^* - \mu\| + \|\theta\|\|\mu^* - \mu\| \leq 2L\|\mu^* - \mu\| \tag{40}$$

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 bound applies to the log-likelihood.

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

$$\mathcal{B}_{A^*}(\mu^*;\mu) \leq \frac{L}{2}\|\mu^* - \mu\|^2 \tag{41}$$

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

ℓ^2-norm Analysis Let us make these statements more precise. A bound on the variance becomes a bound on the variance of the average

$$\operatorname{Var} T(X) = \mathbb{E}\|T(X) - \mu^*\|^2 = \sigma^2 \tag{42}$$

$$\implies \mathbb{E}\left\|\mu^* - \frac{1}{n}\sum_i T(x_i)\right\|^2 = \frac{\sigma^2}{n} \tag{43}$$

eg the variance of the mean is n times smaller than the variance of the samples. Adding a reference mean μ_0 to get the MAP yields

$$\mathbb{E}\left\|\mu^* - \frac{n_0\mu_0 + \sum_i T(x_i)}{n_0 + n}\right\|^2 = \frac{n}{(n+n_0)^2}\sigma^2 + \frac{n_0^2}{(n+n_0)^2}\|\mu^* - \mu_0\|^2 \tag{44}$$

$$= O\left(\frac{\sigma^2}{n}\right) + O\left(\frac{\|\mu^* - \mu_0\|^2}{n^2}\right) \tag{45}$$

so we have a variance term in $O(n^{-1})$ and a bias term decreasing as $O(n^{-2})$. Let's see if we can get similar estimates for arbitrary exponential families !

5 Asymptotic Behaviour

In the limit, the MAP estimates reach a rate $O(1/2n)$. This is shown by approximating the Bregman divergence using a second order Taylor expansion

$$\begin{aligned}\mathcal{B}_{A^*}(\mu^*;\mu) &= \frac{1}{2}(\mu-\mu^*)^\top \nabla^2 A^*(\mu^*)(\mu-\mu^*) + O(\|\mu-\mu^*\|^3) && (46)\\ &= \frac{1}{2}\|\mu^*-\mu\|^2_{\nabla^2 A^*(\mu^*)} + O(\|\mu-\mu^*\|^3) && (47)\end{aligned}$$

where we introduced the Mahalanobis distance induced by the matrix

$$\nabla^2 A^*(\mu^*) = \nabla^2 A(\theta^*)^{-1} = \mathrm{Cov}_{\theta^*}[T(X)]^{-1} =: \boldsymbol{\Sigma}^{-1}\ . \tag{48}$$

To exploit this approximation, let us extend the ℓ^2 norm results to a Mahalanobis distance induced by matrix $\boldsymbol{M}$

$$\mathbb{E}\,\frac{1}{2}\left\|\mu^* - \frac{1}{n}\sum_i T(x_i)\right\|^2_{\boldsymbol{M}} = \frac{1}{2n}\,\mathrm{Tr}(\boldsymbol{M\Sigma}) \tag{49}$$

where $\boldsymbol{\Sigma}$ is the covariance of $T(X)$. We retrieve ${}^{\sigma^2}\!/_n$, the variance divided by the number of samples, when the metric is the identity $\boldsymbol{M}=\boldsymbol{I}$. For the MLE we get

$$\mathbb{E}\,\mathcal{B}_{A^*}\left(\mathbb{E}[T(X)];\frac{1}{n}\sum_i T(X_i)\right) = \frac{d}{2n} + O(n^{-\frac{3}{2}}) \tag{50}$$

and for the MAP we get

$$\begin{aligned}\mathbb{E}\,\frac{1}{2}\left\|\mu^* - \frac{n_0\mu_0+\sum_i T(x_i)}{n_0+n}\right\|^2_{\boldsymbol{\Sigma}^{-1}} &= \frac{nd}{2(n+n_0)^2} + \frac{n_0^2}{(n+n_0)^2}\frac{1}{2}\|\mu^*-\mu_0\|^2_{\boldsymbol{\Sigma}^{-1}} && (51)\\ &= \frac{d}{2n} + O\left(\frac{1+\|\mu^*-\mu_0\|^2_{\boldsymbol{\Sigma}^{-1}}}{n^2}\right)\ . && (52)\end{aligned}$$

or

$$\mathbb{E}\,\mathcal{B}_{A^*}\left(\mathbb{E}[T(X)];\frac{n_0\mu_0+\sum_i T(x_i)}{n_0+n}\right) = \frac{d}{2n} + O(n^{-\frac{3}{2}}) \tag{53}$$

Remark how the variance does not even appear, only the dimension divided by n really matters. I find this result quite spectacular : the convergence speed of MLE is not affected by the covariance of sufficient statistics, and for MAP it matters only in the $O(n^{-2})$ term. Indeed this is because it is hidden within the Bregman divergence loss. Anyway that's all good for asymptotic results, but we are interested in a finite sample analysis. How do we get this ?

6 Bias-Variance Decomposition

6.1 Dual Expectation Pivot.

Let $\hat\mu_n = \frac{n_0\mu_0+\sum_i T(X_i)}{n_0+n}$ be the MAP estimate of the mean (dual) parameter, $\hat\theta_n = \nabla A^*(\hat\mu_n)$ the corresponding natural (primal) parameter, and $\mu_n = \mathbb{E}\left[\hat\mu_n\right] = \frac{n_0\mu_0+n\mu^*}{n_0+n}$ the expected MAP dual estimate Taking μ_n as a pivot, we can decompose the expected suboptimality in 3 terms – e.g.

$$\begin{aligned}\mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*;\hat\mu_n)\right] &= A^*(\mu^*) \pm A^*(\mu_n) - \mathbb{E}\left[A^*(\hat\mu_n)\right] + \mathbb{E}[\langle\hat\theta_n;\hat\mu_n \pm \mu_n - \mu^*\rangle]\\ &= A^*(\mu^*) - A^*(\mu_n) + \mathbb{E}[\mathcal{B}_{A^*}(\mu_n,\hat\mu_n)] + \langle\mathbb{E}[\hat\theta_n] \pm \nabla A^*(\mu_n);\mu_n-\mu^*\rangle\\ &= \underbrace{\mathcal{B}_{A^*}(\mu^*;\mu_n)}_{\text{bias}} + \frac{n_0}{n+n_0}\langle\underbrace{\mu_0-\mu_*}_{\text{bias}};\underbrace{\mathbb{E}[\hat\theta_n]-\theta_n}_{\text{variance}}\rangle + \underbrace{\mathbb{E}[\mathcal{B}_{A^*}(\mu_n,\hat\mu_n)]}_{\text{variance}}\ . && (54)\end{aligned}$$

Remark that unless A^* is ℓ^2, or unless we are in a degenerate case with $\hat\mu_n$ a Dirac, the dual and primal expectation represent different distributions $\theta_n \neq \mathbb{E}\left[\hat\theta_n\right]$, which is why the mixed term (bias-variance scalar product) is non-zero. Asymptotically, these terms behave like : bias $O(n^{-2}$, mixed $O(n^{-3})$ and variance $O(n^{-1})$, but the story is different for finite samples. We know for sure that the bias and the variance term are positive, but the mixed bias-variance term may be negative, which removes some appeal of this decomposition. Fortunately there is another decomposition revolving around the primal expectation.

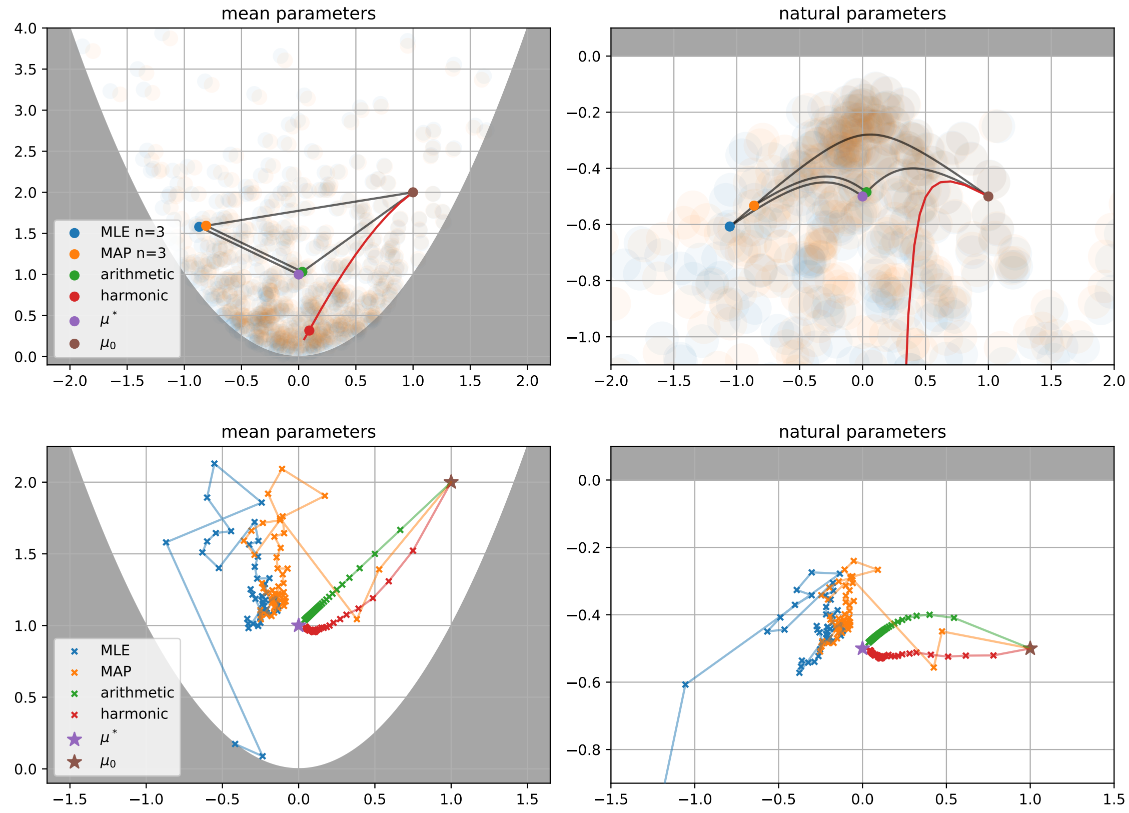

Figure 5: A numerical illustration of the different characters featured in the bias-variance decomposition, for a 1D Gaussian $\mathcal{N}(\mu, \sigma^2)$, for 1 time step and many repetitions with $n = 3$ (top) illustrating the effect of n_0 with the Thales triangle. Or 50 time-steps at once and 1 repetition (bottom), illustrating the convergence behavior.

6.2 Primal Expectation Pivot

Let $\tilde{\theta}_n := \mathbb{E}\left[\hat{\theta}_n\right]$ and $\tilde{\mu}_n = \nabla A(\tilde{\theta}_n)$ be the primal and dual parameters of the primal expectation of the MAP. As described by Pfau (2013, Theorem 0.1), the expected Bregman decomposes like

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*; \hat{\mu}_n)\right] = \underbrace{\mathcal{B}_{A^*}(\mu^*; \tilde{\mu}_n)}_{\text{bias}} + \underbrace{\mathbb{E}\left[\mathcal{B}_{A^*}(\tilde{\mu}_n; \hat{\mu}_n)\right]}_{\text{variance}} \tag{55}$$

$$\mathbb{E}\left[\mathcal{B}_A(\hat{\theta}_n; \theta^*)\right] = \underbrace{\mathcal{B}_A(\mathbb{E}[\hat{\theta}_n]; \theta^*)}_{\text{bias}} + \underbrace{\mathbb{E}\left[\mathcal{B}_A(\hat{\theta}_n; \mathbb{E}[\hat{\theta}_n])\right]}_{\text{variance}} . \tag{56}$$

Illustrations. We show these decompositions for $\mathcal{N}(0, \sigma^2)$ in Figure 6 and $\mathcal{N}(\mu, \sigma^2)$ in Figure 7. In particular for $\mathcal{N}(\mu, \sigma^2)$, we illustrate the characters featured in both of these decompositions $\hat{\mu}_n^{\mathrm{MLE}}, \hat{\mu}_n^{\mathrm{MAP}}, \mu_n, \tilde{\mu}_n, \mu^*$ and μ_0 (and corresponding primal parameters) in Figure 5.

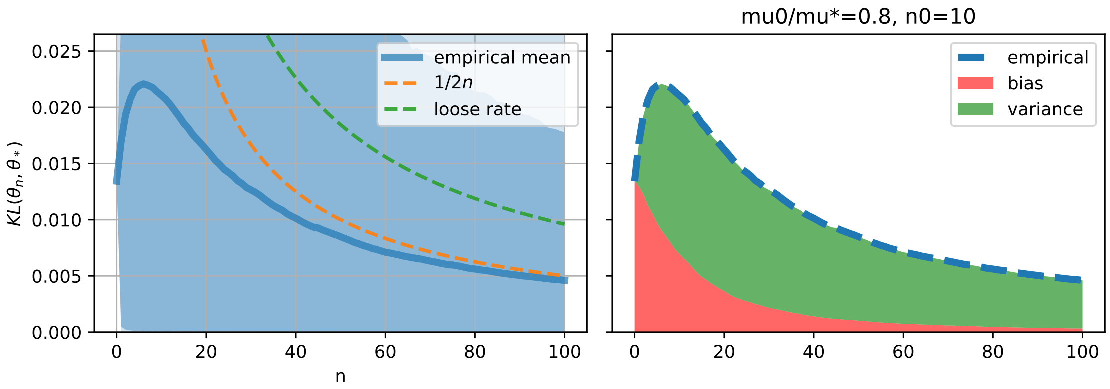

Figure 6: **Gaussian variance $\mathcal{N}(0, \sigma^2)$ example. Left:** training curves and analytic upper bound. **Center:** bias-mixed-variance decomposition, using the arithmetic mean. **Right:** bias-variance decomposition, using the harmonic mean.

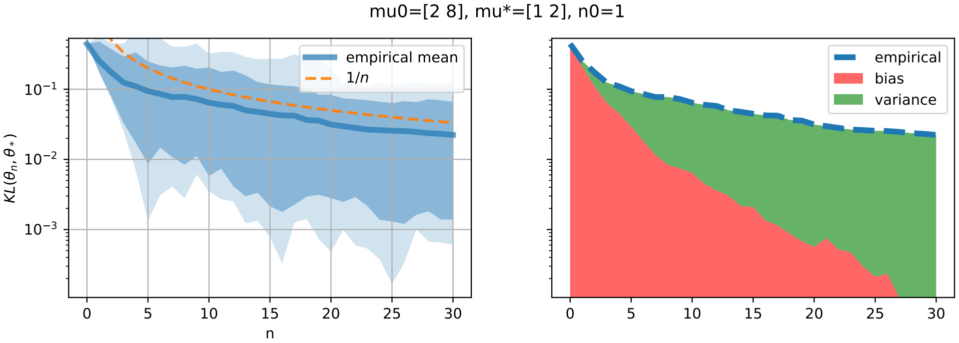

Figure 7: **Full gaussian $\mathcal{N}(\mu, \sigma^2)$ example. Left:** training curves and analytic upper bound. **Center:** bias-mixed-variance decomposition, using the arithmetic mean. **Right:** bias-variance decomposition, using the generalized mean.

6.3 About Variance and Covariance

Lemma 6.1. *Taking $\hat{\theta}_n$ and $\hat{\mu}_n$ as the primal and dual parameters of the MAP estimate, and $\tilde{\mu}_n$ the primal expectation and μ_n the dual expectation, the following identities hold*

$$\begin{aligned} \mathrm{Cov}(\hat{\theta}_n; \hat{\mu}_n) &= \mathbb{E}\left[\mathcal{B}_{A^*}(\hat{\mu}_n; \mu_n) + \mathcal{B}_{A^*}(\mu_n; \hat{\mu}_n)\right] && (57) \\ &= \mathbb{E}\left[\mathcal{B}_{A^*}(\hat{\mu}_n; \tilde{\mu}_n) + \mathcal{B}_{A^*}(\tilde{\mu}_n; \hat{\mu}_n)\right] \ . && (58) \end{aligned}$$

Proof. By the three point identity (or plain derivation), we have

$$\mathcal{B}_{A^*}(\hat{\mu}_n; \mu_n) + \mathcal{B}_{A^*}(\mu_n; \hat{\mu}_n) = \langle \nabla A^*(\hat{\mu}_n) - \nabla A^*(\mu_n); \hat{\mu}_n - \mu_n \rangle = \langle \hat{\theta}_n - \theta_n; \hat{\mu}_n - \mu_n \rangle \tag{59}$$

Taking the expectation, the constant θ_n disappears since $\mathbb{E}\left[\hat{\mu}_n - \mu_n\right] = 0$.

$$\mathbb{E}\left[\langle \hat{\theta}_n - \theta_n; \hat{\mu}_n - \mu_n \rangle\right] = \mathbb{E}\left[\langle \hat{\theta}_n; \hat{\mu}_n \rangle\right] - \langle \mathbb{E}\left[\hat{\theta}_n\right]; \hat{\mu}_n \rangle = \mathrm{Cov}(\hat{\theta}_n; \hat{\mu}_n) \tag{60}$$

which is the first identity. We get the second identity by symmetry between the primal and the dual. □

Note that, as a corollary, the covariance between primal and dual MAP $\mathrm{Cov}(\hat{\theta}_n; \hat{\mu}_n)$ is positive, and it can be seen as a non-linear variance of $\hat{\mu}_n$ – e.g. we can define a variance operator with a non-linear monotone operator ∇A^* plugged in the middle $\mathrm{Var}_{\nabla A^*}(\hat{\mu}_n) := \mathrm{Cov}(\nabla A^*(\hat{\mu}_n); \hat{\mu}_n) \geq 0$.

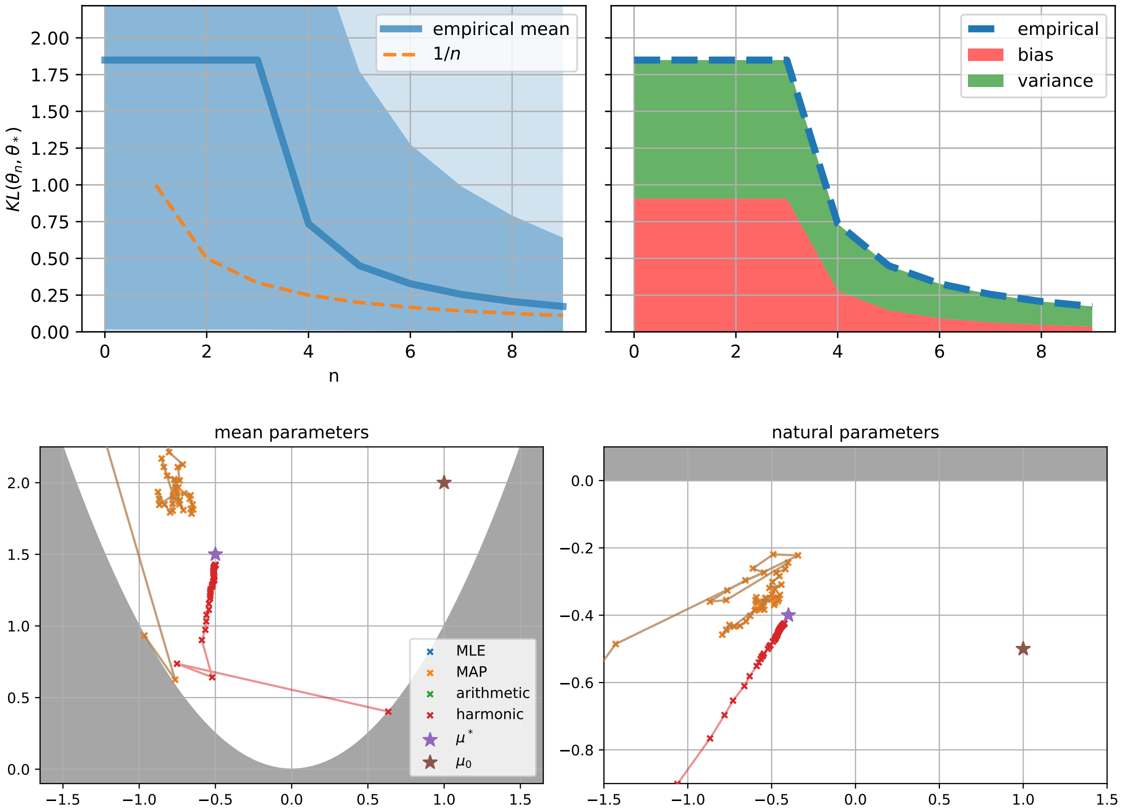

Figure 8: The MLE is biased in primal space.

Expansion We can expand $\hat{\mu}_n - \mu_n$ to get another formulation of this covariance

$$\mathbb{E}\left[\langle\hat{\theta}_n; \hat{\mu}_n - \mu_n\rangle\right] = \mathbb{E}\left[\langle\hat{\theta}_n; \frac{\sum_i T(X_i) - \mu^*}{n + n_0}\rangle\right] \tag{61}$$

$$= \frac{n}{n+n_0}\mathbb{E}\left[\langle\hat{\theta}_n; T(X_0) - \mu^*\rangle\right] \tag{62}$$

$$= \frac{n}{n+n_0}\operatorname{Cov}(\hat{\theta}_n; T(X_0)) \tag{63}$$

where we used the property that all data points play a symmetric role to remove the sum over i.

6.4 Triangle Scaling Exponents

Hanzely et al. (2021) defines the Triangle Scaling Exponent, to characterize an acceleration Bregman methods. The exponent γ of a Bregman divergence $\mathcal{B}$ is defined as the largest coefficient such that $\forall x_0, x_1, x_2, \forall \lambda \in [0,1]$,

$$\mathcal{B}((1-\lambda)x_0 + \lambda x_1; (1-\lambda)x_0 + \lambda x_2) \leq \lambda^\gamma \mathcal{B}(x_1; x_2)\ . \tag{64}$$

Geometrically, for any triangle, it tells us a shape within which the geodesic triangle is contained (to clarify with a picture.) It is related to Thales triangle. We precisely have a Thales triangle with the dual expectation pivot, meaning we can relate the MAP variance term with the MLE expected suboptimality. Writing $\lambda = \frac{n}{n+n_0}$, we observe

$$\hat{\mu}_n^{\text{MAP}} = (1-\lambda)\mu_0 + \lambda\hat{\mu}_n^{\text{MLE}} \tag{65}$$

$$\mu_n = (1-\lambda)\mu_0 + \lambda\mu^* \tag{66}$$

thus

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_n; \hat{\mu}_n^{\text{MAP}})\right] \leq \left(\frac{n}{n+n_0}\right)^\gamma \mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*; \hat{\mu}_n^{\text{MLE}})\right]\ . \tag{67}$$

This relationship between the MAP variance and the MLE may be void, as the MLE expected loss may be infinite. But in some settings it might be helpful to characterize the MAP convergence. For instance, this same paper says that for a gaussian variance, $\gamma = 1$. Then they proceed to define an intrinsic scaling exponent, with a local scaling factor. This intrinsic exponent is always 2 for Bregman divergences induced by a convex and $\mathcal{C}^2$ function. They use this fact to derive an accelerated algorithm.

7 Optimization Perspective

Problem. We want to solve the following problem

$$f^* = \min_\theta \mathbb{E}[f(\theta;X)] := -\langle \mathbb{E}[T(X)]; \theta\rangle + A(\theta) \tag{68}$$

where A is a convex function. We assume access to stochastic oracles $X \sim p(X|\theta^*)$ – e.g. we observe X such that $\mathbb{E}[T(X)] = \mu^*$ and we can compute stochastic function values and gradients

$$f(\theta;X) = -\langle T(X); \theta\rangle + A(\theta) \tag{69}$$

$$\nabla f(\theta;X) = \nabla A(\theta) - T(X)\ . \tag{70}$$

It is important to note that our problem is merely a stochastic linear tilt of the convex function A, yet none of the convergence rates we studied applied to this setting.

Algorithms. We observe n data points X_i. We are studying two algorithms solving this problem : MLE and MAP.

$$\hat\mu_n^{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n T(X_i) \qquad \hat\theta_n^{\text{MLE}} = \nabla A^{-1}(\hat\mu_n^{\text{MLE}}) \tag{71}$$

$$\hat\mu_n^{\text{MAP}} = \frac{n_0\mu_0 + \sum_{i=1}^n T(X_i)}{n_0+n} \qquad \hat\theta_n^{\text{MAP}} = \nabla A^{-1}(\hat\mu_n^{\text{MAP}}) \tag{72}$$

Interpretation. As we will see in the two upcoming sections, MAP can be seen as an application of more generic iterative algorithms with divergence $\mathcal{B}_A$ and initialization θ_0 :

- Stochastic Bregman Proximal Point with learning rate $\frac{1}{n_0+n}$,
- Stochastic Mirror Descent with learning rate $\frac{1}{n_0+n+1}$.

In both cases, $f(\theta)$ is convex and 1-smooth and 1-strongly convex relative to the potential A. Be cautious that the learning rate is decreasing.

Goal. We want to bound the suboptimality $f(\theta) - f^*$ as a function of the number of samples.

8 Stochastic Bregman Proximal Point

8.1 Correspondence with MAP

For a dataset of independent samples $\mathcal{D}_n = \{X_1, \dots, X_n\}$, the posterior can be seen iteratively

$$p(\theta|\mathcal{D}_n) = p(\theta|X_n, \mathcal{D}_{n-1}) \propto p(X_n|\theta, \mathcal{D}_{n-1})p(\theta|\mathcal{D}_{n-1}) = p(X_n|\theta)p(\theta|\mathcal{D}_{n-1})\ . \tag{73}$$

For the MAP of the exponential family with a Bregman conjugate prior, the posterior takes the form

$$\begin{aligned}
-\log p(\theta|\mathcal{D}_n) &= \sum_{i=1}^n -\log p(X_i|\theta) - \log p(\theta) + \text{cst} && (74)\\
&= \sum_i -\langle T(X_i); \theta\rangle + nA(\theta) + n_0\mathcal{B}_A(\theta;\theta_0) + \text{cst} && (75)\\
&= (n+n_0)A(\theta) - \langle n_0\mu_0 + \sum_i T(X_i); \theta\rangle + \text{cst} && (76)\\
&= (n+n_0)\mathcal{B}_A(\theta; \hat\theta_n^{\text{MAP}}) + \text{cst}\ , && (77)
\end{aligned}$$

so we immediately get the recursion

$$\begin{aligned}
\hat\theta_n^{\text{MAP}} &= \operatorname*{argmin}_\theta -\log p(\theta|\mathcal{D}_n) && (78)\\
&= \operatorname*{argmin}_\theta -\log p(X_n|\theta) - \log p(\theta|\mathcal{D}_{n-1}) && (79)\\
&= \operatorname*{argmin}_\theta f(\theta;X_n) + (n+n_0)\mathcal{B}_A(\theta;\hat\theta_{n-1}^{\text{MAP}})\ . && (80)
\end{aligned}$$

Coincidentally, this is exactly the formula of a Stochastic Bregman Proximal Point update on the function $f(\theta) = \mathbb{E}_X[-\log p(X|\theta)]$, with step-size $\gamma_n = \frac{1}{n+n_0}$. As a reminder this algorithm is defined with the update

$$\hat{\theta}_n = \operatorname*{argmin}_\theta \gamma_n f(\theta; X_n) + \mathcal{B}_A(\theta; \hat{\theta}_{n-1}) \tag{81}$$

$$\iff \nabla A(\theta_{n-1}) = \nabla A(\theta_n) + \gamma_t \nabla f(\theta_n; X_t) \ . \tag{82}$$

In words, if it converges to θ^*, it can be seen as a backward mirror ascent starting from θ^*.

8.2 Straightforward Analysis

8.2.1 Deterministic Bregman Proximal Point

Deterministic Bregman proximal point. was first published in Eckstein (1993) with an asymptotic convergence property, soon completed by a finite sample convergence rate in Chen and Teboulle (1993). This convergence rate relies on the three points lemma (Appendix B)

$$\forall \theta, \gamma_t f(\theta) + \mathcal{B}(\theta; \theta_t) \geq \gamma_t f(\theta_{t+1}) + \mathcal{B}(\theta; \theta_{t+1}) + \mathcal{B}(\theta_{t+1}; \theta_t) \ . \tag{83}$$

Monotonicity of Function Values. Applying it with $\theta = \theta_t$, we prove that function values are decreasing

$$f(\theta_{t+1}) \leq f(\theta_t) - \frac{1}{\gamma_t} \underbrace{(\mathcal{B}(\theta_{t+1}; \theta_t) + \mathcal{B}(\theta_t; \theta_{t+1}))}_{=:\mathcal{S}(\theta_t;\theta_{t+1})} \tag{84}$$

where we introduced $\mathcal{S}$ the symmetrized Bregman divergence

$$\mathcal{S}(x; y) = \mathcal{B}(x; y) + \mathcal{B}(y; x) = \langle \nabla A(x) - \nabla A(y); x - y \rangle \ . \tag{85}$$

Descent Lemma and Telescopic Sum. Applying the three point lemma with $\theta = \theta^*$ (assuming such a minimizer exists), we obtain a telescopic sum

$$\gamma_t(f(\theta_{t+1}) - f(\theta^*)) + \mathcal{B}(\theta_{t+1}; \theta_t) \leq \mathcal{B}(\theta^*; \theta_t) - \mathcal{B}(\theta^*; \theta_{t+1}) \tag{86}$$

$$\implies \sum_{t=0}^{T-1} \gamma_t(f(\theta_{t+1}) - f(\theta^*)) + \sum_{t=0}^{T-1} \mathcal{B}(\theta_{t+1}; \theta_t) \leq \mathcal{B}(\theta^*; \theta_0) - \mathcal{B}(\theta^*; \theta_T) \leq \mathcal{B}(\theta^*; \theta_0) \ , \tag{87}$$

which proves on the way that the total path length $\sum_{t=0}^{t-1} \mathcal{B}(\theta_{t+1}; \theta_t)$ is bounded. Using the monotonicity of $f(\theta_t)$, we get

$$f(\theta_T) - f(\theta^*) \leq \frac{\mathcal{B}(\theta^*; \theta_0)}{\sum_{t=0}^{T-1} \gamma_t} \ . \tag{88}$$

This upper bound converges to 0 whenever the sum of step-sizes converges to 0. This rate can be arbitrarily fast because the oracle is strong. Indeed, taking an inifinite step-size solves the problem with one iteration, but the update is (theoretically) harder to solve for larger step-sizes.

8.2.2 Stochastic Bregman Proximal Point.

While Eckstein (1998) shows some asymptotic convergence of the Bregman proximal point algorithm when each iteration is only solved approximately, we are not aware of published analysis when iterations are stochastic.

Variance Assumption. The deterministic proof does not transpose immediately to the stochastic setting. We need an assumption on the quality of the stochastic estimates $f(.;x)$. As we will see soon, the most straightforward such assumption is

$$\boxed{\mathbb{E}\left[f(\theta_{t+1}) - f(\theta_{t+1}; X_{t+1})\right] \leq \gamma_t \sigma^2 \ ,} \tag{89}$$

e.g. we relate the stochastic function value on the next iterate to the function value. Asymptotically, this assumption is equivalent to the one of Hanzely and Richtárik (2018). To see this, linearize $f(\theta_{t+1})$ and $f(\theta_{t+1}; X_{t+1})$ in θ_t

$$\mathbb{E}\left[f(\theta_{t+1}) - f(\theta_{t+1}; X_{t+1})\right] \approx \mathbb{E}\left[f(\theta_t) - f(\theta_t; X_{t+1})\right] + \mathbb{E}\left[\langle \nabla f(\theta_t) - \nabla f(\theta_t; X_{t+1}); \theta_{t+1} - \theta_t \rangle\right] \tag{90}$$

$$= 0 + \mathbb{E}\left[\langle \nabla f(\theta_t) - \nabla f(\theta_t; X_{t+1}); \theta_{t+1} \rangle\right] \tag{91}$$

$$= \mathrm{Cov}(-\nabla f(\theta_t; X_{t+1}); \theta_{t+1}) \ . \tag{92}$$

Monotonicity of Expected Function Values. Applying the tree point lemma with $\theta = \theta_t$ gives

$$f(\theta_{t+1}, X_{t+1}) \leq f(\theta_t, X_{t+1}) - \frac{1}{\gamma_t}\left(\mathcal{B}(\theta_{t+1}; \theta_t) + \mathcal{B}(\theta_t; \theta_{t+1})\right) \tag{93}$$

$$= f(\theta_t, X_{t+1}) - \frac{1}{\gamma_t}\mathcal{S}(\theta_{t+1}; \theta_t) \tag{94}$$

$$\implies \mathbb{E}\left[f(\theta_{t+1}, X_{t+1})\right] \leq f(\theta_t) - \frac{1}{\gamma_t}\mathbb{E}\left[\mathcal{S}(\theta_{t+1}; \theta_t)\right] \ . \tag{95}$$

This inequality bears on $\mathbb{E}\left[f(\theta_{t+1}, X_{t+1})\right]$, which features a correlation between θ_{t+1} and X_{t+1}, and is different from $f(\theta_{t+1})$. Combining (95) with the variance assumption gives

$$\mathbb{E}\left[f(\theta_{t+1})\right] \leq \mathbb{E}\left[f(\theta_{t+1}; X_{t+1})\right] + \gamma_t \sigma^2 \tag{96}$$

$$\leq f(\theta_t) + \gamma_t \sigma^2 - \frac{1}{\gamma_t}\mathbb{E}\left[\mathcal{S}(\theta_{t+1}; \theta_t)\right] \tag{97}$$

from which we see that monotonicity is guaranteed if

$$\sigma^2 \leq \frac{1}{\gamma_t^2}\mathbb{E}\left[\mathcal{S}(\theta_{t+1}; \theta_t)\right] \ . \tag{98}$$

This is quite annoying, as our variance assumption requires σ^2 to be large, and this inequality requires σ^2 to be small – the step we take should be greater than the variance.

Descent Lemma and Telescopic Sum. Similarly, applying the three points lemma with θ^* yields

$$\gamma_t(f(\theta_{t+1}, X_{t+1}) - f(\theta^*)) + \mathcal{B}(\theta_{t+1}; \theta_t) \leq \mathcal{B}(\theta^*; \theta_t) - \mathcal{B}(\theta^*; \theta_{t+1}) \ . \tag{99}$$

Ignoring the distance between iterates $\mathcal{B}(\theta_{t+1}; \theta_t)$, taking the tower expectation, and telescoping the sum, we get

$$\sum_{t=0}^{t-1} \gamma_t \mathbb{E}\left[f(\theta_{t+1}, X_{t+1}) - f(\theta^*)\right] \leq \mathcal{B}(\theta^*; \theta_0) \ . \tag{100}$$

Then, taking the iterate with the least expected loss, we get the rate

$$\min_{1 \leq t \leq T} \mathbb{E}\left[f(\theta_t)\right] - f(\theta^*) \leq \frac{\mathcal{B}(\theta^*; \theta_0)}{\sum_{t=0}^{T-1} \gamma_t} + \frac{\sum_{t=0}^{T-1} \gamma_t^2}{\sum_{t=0}^{T-1} \gamma_t}\sigma^2 \ . \tag{101}$$

which is very analogous to standard rates for SGD. It does converge, whenever $\sum_{t=0}^{T-1} \gamma_t$ diverges to $+\infty$ much faster than $\sum_{t=0}^{T-1} \gamma_t^2$. But in our case, $\gamma_t \propto t^{-1}$, and we merely get a $O(\frac{1}{\log(t)})$ rate. With the sums $\sum \gamma_t$ and $\sum \gamma_t^2$, we would need a $O(\frac{1}{\sqrt{t}})$ step-size to ensure a $O(\frac{1}{\sqrt{t}})$ convergence rate, but then the iterate becomes a weighted average of the sufficient statistics, giving more weight to the end of the tail. This may be suited for online learning, but not for our stationary setting.

Application to the MAP. The variance assumption (89) translates to $\mathrm{Cov}(T(x_n), \theta_{n+1}) \leq \gamma_n \sigma^2$ in the exponential family MAP setting – e.g. exactly the same assumption as in Hanzely and Richtárik (2018).

9 Stochastic Mirror Descent (SMD)

9.1 MAP as SMD

In this section, we show that MAP estimate can also be seen as the iterates of stochastic mirror descent with mirror map A.

Mirror Descent. Let us recall mirror descent iteration

$$\begin{aligned} \theta_{t+1} &:= \operatorname*{argmin}_\theta \gamma \ell_f(\theta; \theta_t) + \mathcal{B}_A(\theta; \theta_t) && (102) \\ &= \nabla A^*(\nabla A(\theta_t) - \gamma \nabla f(\theta_t)) && (103) \end{aligned}$$

where $\ell_f(\theta; \theta_t) = f(\theta_t) + \langle \nabla f(\theta_t), \theta - \theta_t \rangle$ is the linear approximation of f in θ_t evaluated at θ. Solving this problem require solving problems of the form $\operatorname{argmin}_\theta -\langle c, \theta \rangle + A(\theta)$, eg computing the convex conjugate of A. Note that finding θ_* is done with 1 step of mirror descent with step-size 1. Indeed plugging in definitions of f and μ, and assuming constant step-size γ yields

$$\begin{aligned} \mu_{t+1} &= \mu_t - \gamma(\mu_t - \mu_*) && (104) \\ \implies \mu_t &= \mu_* + (1-\gamma)^t (\mu_0 - \mu_*) && (105) \end{aligned}$$

which shows exponential convergence and 1-step convergence when $\gamma = 1$.

Stochastic Mirror Descent Let $f(\theta, x) = -\log p(x|\theta)$. MAP iterations can be cast as stochastic mirror descent (SMD), with $g_t = \nabla f(\theta_t, x_{t+1}) = A(\theta_t) - T(x_{t+1})$ as a stochastic estimate of $\nabla f(\theta_t)$. Resuming from the Stochastic Bregman Proximal Point formula (80), we write $\hat{\theta}_n$ for the n-th MAP iterate, and do some Bregman manipulations

$$\begin{aligned} \hat{\theta}_{n+1} &= \operatorname*{argmin}_\theta -\langle T(x_{t+1}), \theta \rangle + A(\theta) + (n_0 + n)\mathcal{B}_A(\theta; \theta_n) && (106) \\ &= \operatorname*{argmin}_\theta -\langle T(x_{t+1}), \theta \rangle + A(\theta_n) + \langle \nabla A(\theta_n), \theta - \theta_n \rangle + (n_0 + n + 1)\mathcal{B}_A(\theta; \theta_n) && (107) \\ &= \operatorname*{argmin}_\theta \ell_f(\theta; \theta_n, x_{t+1}) + (n_0 + n + 1)\mathcal{B}_A(\theta; \theta_n) && (108) \end{aligned}$$

where $\ell_f(\theta; \theta_n, x_{t+1})$ is the stochastic 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 $\gamma_n = \frac{1}{n_0+n+1}$.

We know of 2 papers that study Stochastic Mirror Descent (SMD) under relative smoothness assumptions – Hanzely and Richtárik (2018) and Dragomir et al. (2021). We are now going to review these works.

9.2 Work by Hanzely

General Approach. Hanzely and Richtárik (2018) got convergence rates in the smooth strongly-convex case with tail averaging. With constant step-size, they proved linear convergence down to a variance ball. With step-size $\gamma_t = n_0 + t$, they proved a rate $\tilde{O}(\frac{1}{t})$. These results match the rates for standard SGD. Let us have a look at their variance assumption (Hanzely and Richtárik, 2018, Assumption 5.1). 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 an upper bound on the covariance between the gradient update $-g_t$ and the next iterate $\theta_{t+1} = \nabla A^*(\nabla A(\theta_t) - \gamma_t g_t)$ that should hold for all time steps

$$\begin{aligned} \gamma_t \sigma^2 &\geq \mathrm{Cov}(-g_t, \theta_{t+1} | X_{1\ldots t}) && (109) \\ &= \mathbb{E}\left[\langle \nabla f(\theta_t) - g_t, \theta_{t+1} \rangle\right] && (110) \end{aligned}$$

where γ_t is the step-size, and expectations are conditional on all of the past. Remark that when we plug in $A^* = \|.\|^2$, we recover the gradient variance typical of SGD. If we introduce $\bar{\theta}_{t+1} = \nabla A^*(\nabla A(\theta_t) - \gamma_t \nabla f(\theta_t))$ the theoretical output of a deterministic Bregman gradient step, then following Lemma 6.1 we can rewrite this covariance as the expectation of symmetrized Bregman between stochastic and deterministic iterates $\mathbb{E}_{X_t}\left[\mathcal{B}_A(\theta_{t+1}; \bar{\theta}_{t+1}) + \mathcal{B}_A(\bar{\theta}_{t+1}; \theta_{t+1})\right]$. This view is useful to understand that we are measuring the variance of each iteration, which is necessarily positive.

MAP Special Case. In our setting,

$$\nabla f(\theta_t) - g_t = T(X_{t+1}) - \mathbb{E}[T(X)] \tag{111}$$

so that σ is really a bound on the covariance of the sufficient statistics $T(X_{t+1})$ with the next iterate θ_{t+1}, which itself depends on $T(X_{t+1})$ via a coefficient γ_t and a non-linearity ∇A^*.

$$\begin{aligned} &\mathrm{Cov}(T(X_{t+1}), \theta_{t+1} | X_{1\ldots t}) \leq \gamma_t \sigma^2 && (112)\\ =\, &\mathrm{Cov}(T(X_{t+1}), \nabla A^* \left(\gamma_t T(X_{t+1}) + (1-\gamma_t)\mu_t\right) | X_{1\ldots t}) \,. && (113)\end{aligned}$$

In other words, we need for all $\gamma_t \in [0; \gamma_0]$, and the subset of μ that can be reached by on the trajectory μ_t,

$$\boxed{\mathrm{Cov}_X\Big(T(X), \nabla A^*(\gamma T(X) + (1-\gamma)\mu)\Big) \leq \gamma\sigma^2\,.} \tag{114}$$

Unfortunately this bound is not trivially satisfied, even for simple exponential family members. For instance, centered gaussians variance estimation $\mathcal{N}(0, \mu^*)$ does not verify (114). Indeed the covariance (114) we want to bound is equal to

$$\frac{1}{2}\mathbb{E}_{X\sim\mathcal{N}(0,\mu^*)}\left[\frac{\mu *^2 - X^2}{X^2\gamma + (1-\gamma)\mu^2}\right] = \frac{1}{\gamma}(1+a)\mathbb{E}_{X\sim\mathcal{N}(0,1)}\left[\frac{1}{a+X^2}\right] - \frac{1}{\gamma} \tag{115}$$

where $a = \frac{1-\gamma}{\gamma}\frac{\mu}{\mu^*}$ When $\sigma_0 \to 0$, as might happen, this value explodes.

Straightforward analysis When we plug this bound into the expected suboptimality formula (27) with the maximum likelihood estimate $\hat{\mu} = \frac{1}{n}\sum_i T(X_i)$, we get

$$\begin{aligned}\mathbb{E}_{X_{1\ldots n}}\left[\mathcal{B}_{A^*}(\mu_*; \hat{\mu})\right] = A^*(\mu_*) &\overbrace{- \mathbb{E}[A^*(\hat{\mu})]}^{\leq -A^*(\mathbb{E}[\hat{\mu}])} \\ &+ \frac{1}{n}\sum_i \mathbb{E}[\underbrace{\mathbb{E}[\langle T(X_i) - \mu_*; \nabla A^*(\hat{\mu})\rangle | X_j, j \neq i]}_{\leq \sigma^2/n (114) \text{ with } \mu = \frac{1}{n-1}\sum_{j\neq i} T(x_j)}] \leq \frac{\sigma^2}{n}\end{aligned} \tag{116}$$

where we used the decomposition $\hat{\mu} = \frac{1}{n}T(X_i) + (1-\frac{1}{n})\frac{1}{n-1}\sum_{j\neq i} T(X_j)$ to apply (114). In words, this variance assumption on $T(X)$ and A^* immediately gives us a bound on the suboptimality.

This is too simple to be true, as shown on our trailing example. In particular we show that it cannot hold for all possible values of μ within the double expectation in (116). The mean parameter μ takes values arbitrarily close from 0, which makes the covariance explode.

Bias-Variance Perspective. As shown in equation (63), covariance (112) is an upper bound on the variance term in either the primal (56) or dual (54) expectation pivot decompositions

$$\mathrm{Cov}(\hat{\mu}_n, \hat{\theta}_n) = \frac{n}{n+n_0}\mathbb{E}_{X_{1:n-1}}\left[\underbrace{\mathrm{Cov}(\hat{\theta}_n; T(X_n) | X_{1:n-1})}_{\leq \sigma^2/n (114)}\right] \leq \frac{n\sigma^2}{(n+n_0)^2} \tag{117}$$

so we recover a $O(n^{-1})$ upper bound on the variance term and we are still looking for the $O(n^{-2})$ rate for the bias term. It is simpler to consider the primal pivot, but still hard to say anything about $\mathcal{B}_{A^*}(\mu^*; \tilde{\mu}_n) = \mathcal{B}_A(\mathbb{E}[\hat{\theta}_n]; \theta^*)$ without further assumptions.

9.3 Work by the French Musketeers

Dragomir et al. (2021) use a quite different assumption, bearing on a Bregman measure of the stochastic gradients norm at the optimum

$$2\gamma^2\sigma^2 \geq \mathbb{E}_X\left[\mathcal{B}_{A^*}(\mu_t - 2\gamma_t \nabla f(\theta^*, X); \mu_t)\right] \ . \tag{118}$$

When A is strongly convex, this hypothesis is akin to bounding the variance of the gradient at the optimum, whereas (109) is akin to bounding gradient variance everywhere. This comparison makes this assumption more interesting, as it might apply to unbounded strongly convex function, like modern analysis of SGD.

However it does use a metric which depends on the trajectory μ_t. It is the flaw of both these variance assumptions (109) and (118), as well the proximal point variance assumption (89), that they depend on the stochastic trajectory. Indeed this trajectory has a non-zero probability of being arbitrarily close from the border of the domain, where $\mathcal{B}_A$ may get extremely ill-conditioned.

Their work can be seen as a Bregman generalization of Gower et al. (2019).

9.4 Work by Loizou

An assumption pointed out by Nicolas Loizou is

$$f(\theta^*) - \mathbb{E}_{X\sim\theta^*}[\min_\theta f(\theta; X)] < \infty \tag{119}$$

for which a sufficient condition is

$$\forall X, \min_\theta f(\theta; X) > -\infty \iff \max_\theta p(X|\theta) < \infty \tag{120}$$

which holds in a lot of settings but not all. For instance, a gaussian $\mathcal{N}(\mu, \sigma^2)$ can overfit on a single data point and give it infinite density – eg converge to a Dirac. However we can overcome this issue by grouping data points : a gaussian can not overfit on two points. Aggregating points is simple : average sufficient statistics, exactly like in MLE or MAP $T'(X_1, X_2) = \frac{T(X_1)+T(X_2)}{2}$. So in general we may say that this hypothesis holds, at the expense of dividing the number of samples by some number n/k. As mentioned earlier, though, this assumption does not convert to a $O(t^{-1})$ rate when the step-size is itself $O(t^{-1})$.

10 Discussion and other idea

There are two main bottlenecks when casting MAP as an iterative algorithm.

1. The first one is the variance definition which always end up being infinite when asked to be uniform on every possible trajectories, if the log-partition has some barrier structure. Except for D'Orazio et al. (2021) which holds in many cases.

2. The second one is the step-size scheduling. $O(\frac{1}{t})$ is decreasing too fast for most proof techniques, except Dragomir et al. (2021). Even in the Euclidean setting where the variance does not matter, it results in a disappointing rate $O(\frac{1}{\log t})$.

Regarding this second point, it might be possible to do something comparable to Gower et al. (2019) or Dragomir et al. (2021) who get a $O(\frac{1}{t})$ rate with a $O(\frac{1}{t})$ step-size. Take m and L as the (relative) strong-convexity and smoothness constants of the problem, along with $K = \frac{L}{m}$ the condition number.Their descent lemma looks like

$$B_{t+1} \leq (1 - \gamma_t m)B_t + \gamma_t^2\sigma^2 \tag{121}$$

where B_t is some measure of suboptimality, in this case the ℓ^2 distance between θ^t and θ^* ; and $\gamma_t \leq \frac{1}{2L}$. On the other hand, the problem of most Bregman analysis (Hanzely and Richtárik, 2018; D'Orazio et al., 2021) is that they do not get this γ^2 in front the variance. Instead they simply get a γ, because they do not benefit from the homogeneity of the norm

$$B_{t+1} \leq (1 - \gamma_t m)B_t + \gamma_t\sigma^2 \ . \tag{122}$$

The problem is that this exponent 2 is crucial to get fast convergence with a $O(\frac{1}{t})$ step-size. Otherwise the variance is too large and we need to decrease the step-size more slowly to reach the optimum quickly. Now our question is what happens if we get something in-between ? e.g.

$$B_{t+1} \leq (1 - \gamma_t m) B_t + \gamma_t^\alpha \sigma^2 \ . \tag{123}$$

where $1 < \alpha < 2$ is some exponent characterizing the geometry of the problem.

Step-size. Inspired by Gower et al. (2019) we set

$$\gamma_t := \frac{1}{m}\left(1 - \left(\frac{t}{t+1}\right)^\alpha\right) \ . \tag{124}$$

The function $x \mapsto (1+x)^\beta$ is convex, meaning that

$$\left(\frac{t}{t+1}\right)^\alpha = \left(1 - \frac{1}{t+1}\right)^\alpha \gtrsim 1 - \frac{\alpha}{t+1} \tag{125}$$

where the $\gtrsim$ symbol stands for greater or equal $\geq$ and asymptotically equivalent $\sim$. Back to the step-size, we get

$$\gamma_t \lesssim \frac{\alpha}{m(t+1)} \ . \tag{126}$$

From this we draw two conclusions. First the step-size is in the right asymptotic category $\gamma_t \in \Theta(t^{-1})$. Second, we have a sufficient condition for the step-size to be small enough $\gamma_t \leq \frac{1}{2L}$ and the descent lemma to apply

$$\gamma_t \leq \frac{1}{2L} \impliedby \frac{\alpha}{m(t+1)} \leq \frac{1}{2L} \iff \frac{2\alpha L}{m} \leq t + 1 \impliedby t \geq t_0 := \lceil 2\alpha K \rceil - 1 \tag{127}$$

where $K = L/m$ is the condition number of the problem. Consequently, when $t < t_0 :== \lceil 2\alpha K \rceil$, we set $\gamma_t = \frac{1}{2L}$. Let us correct our definition to account for this revision

$$\gamma_t := \begin{cases} \frac{1}{2L} & \text{if } \ t < t_0 = \lceil 2\alpha K \rceil - 1, \\ \frac{1}{m}\left(1 - \left(\frac{t}{t+1}\right)^\alpha\right) & \text{otherwise.} \end{cases} \tag{128}$$

Telescopic Sum. For $t \geq t_0$, the step-size definition means that the decreasing factor of the descent lemma is

$$1 - \gamma_t m = \left(\frac{t}{t+1}\right)^\alpha \ . \tag{129}$$

Multiplying both sides of the descent lemma with $(t+1)^\alpha$ gives

$$(t+1)^\alpha B_{t+1} \leq t^\alpha B_t + ((t+1)\gamma_t)^\alpha \sigma^2 \ . \tag{130}$$

Thanks to (126), we know that $(t+1)\gamma_t \leq \frac{\alpha}{m}$, thus the second term or the upper bound is upper bounded as

$$((t+1)\gamma_t)^\alpha \sigma^2 \leq \left(\frac{\alpha}{m}\right)^\alpha \sigma^2 \ . \tag{131}$$

By recursion (or telescopic summing), we get

$$t^\alpha B_t \leq t_0^\alpha B_{t_0} + (t - t_0)\left(\frac{\alpha}{m}\right)^\alpha \sigma^2 \ . \tag{132}$$

It is possible to get the constant step-size rate for this B_{t_0}, but for now, we already see that

$$B_t \leq \left(\frac{t_0}{t}\right)^\alpha B_{t_0} + \frac{t - t_0}{t^\alpha}\left(\frac{\alpha}{m}\right)^\alpha \sigma^2 \in \Theta(t^{1-\alpha}) \tag{133}$$

where $-1 < 1 - \alpha < 0$ and this slowly decreasing factor is coming from the variance term.

11 Related Work

Exponential families are a mainstream tool in machine learning. They are used to generalize linear regression (McCullagh and Nelder, 1989), PCA (Collins et al., 2001) or k-means (Banerjee et al., 2005) to diverse data types and distributions. Wainwright and Jordan (2008, Chapter 3) give a great overview of exponential families and their dual structure.

Agarwal and Daumé (2010) highlight that the MAP problem of exponential families is also a Bregman median problem. They then use this equivalence to justify the use of conjugate prior in Bayesian estimation and hybrid generative-discriminative modeling. Note that this equivalence holds only if sufficient statistics are within $\operatorname{Dom} A^* = \operatorname{Im}(\nabla A)$, a point ignored by authors yet critical as it does not hold for multivariate normals with unknown covariance. Raskutti and Mukherjee (2015) highlight the equivalence between mirror descent in θ and natural gradient descent in μ (or vice-versa) for the exponential family. Then they use the optimality result by Shunichi Amari on Natural Gradient Descent to show that Mirror Descent iterates reach the Cramer-Rao lower bound.

Kunstner et al. (2021) show with an exquisite elegance that Expectation-Maximization in exponential families can be cast as mirror descent on the negative log-likelihood. Then they use recent mirror descent rates based on the idea of *relative smoothness* (Birnbaum et al., 2011; Bauschke et al., 2017; Lu et al., 2018) to show convergence of EM in KL-divergence, including for the ubiquitous Gaussian Mixture model, which had resisted bounding attempts for 50 years.

Relative smoothness gave rise to sibling hypothesis, such as *relative continuity* (Lu, 2019), which was used to characterize non-differentiable functions – eg Hinge loss. These functions are typically assumed to be Lipschitz, but that is not always true – eg squared Hinge loss. Relative continuity generalizes Lipschitz-ness by upper bounding the gradient norm (or expected norm of stochastic gradients) with the ratio between a Bregman and the ℓ^2 distance. Although this hypothesis is far from intuitive, Lu (2019) obtained convergence rates for mirror descent that are very similar to typical SGD analysis. In this work we are concerned with strictly convex differentiable functions $A(\theta)$, to ensure a clean bijection between natural and mean parameters. Consequently, we do not further investigate relative continuity.

Pfau (2013) gives a bias variance decomposition for Bregman divergences. However it does not give interesting conclusions for our problem of interest.

Bubeck and Eldan (2015) shows that the entropy of an exponential family defined on a compact support $\mathcal{X}$ is a self-concordant barrier on this compact set. They extend the definition of self-concordance to ν-self-concordance, stating that the gradient should be bounded by the square root of the hessian along any direction. In 1 dimension, their result also means that the log-partition is self-concordant as well. In general, does this mean that the log-partition is self-concordant (e.g. is there a link between the dual function on a line and the line of the primal , probably not)?

In the realm of self-concordance, Ostrovskii and Bach (2021) shows a finite sample convergence rate for the MLE (or other M-estimators) when the log-likelihood is self-concordant. This convergence rate holds only after a given number of samples have been seen, and they give an innteresting formula for this number. To get this results, they use the fact that self-concordant losses are upper bounded by a quadratic in a neighborhood of the optimum.

TODO 3 threads of literature search: optimization, statistics, learning theory. This last one is not covered yet. It is more interested in finite sample results than statistics. Contact Alexander Rakhlin

Bregman proximal point

high probability bounds

11.1 Alternative Research Tracks

- A picture / geometric understanding of the equality between ELBO (Jensen) and relative smoothness for EM.
- Bregman is about general geometry, can we bring it to the non-convex realm via local analysis ?

- Can we design an algorithm with higher order Bregman divergences ? Does it mean anything ? The Bregman is a quadratic approximation of a function, yielding a geometry. Higher order does not carry such meaning. But we would like to involve some quadratic function information within mirror descent !

References

Agarwal, A. and Daumé, H. (2010). A geometric view of conjugate priors. *Machine learning*, 81(1):99–113.

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. (2005). Clustering with bregman divergences. *JMLR*, 6(10).

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.

Bubeck, S. and Eldan, R. (2015). The entropic barrier: a simple and optimal universal self-concordant barrier. *COLT*.

Chen, G. and Teboulle, M. (1993). Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543.

Collins, M., Dasgupta, S., and Schapire, R. E. (2001). A generalization of principal components analysis to the exponential family. In *NeurIPS*, volume 13, page 23.

Dragomir, R.-A., Even, M., and Hendrikx, H. (2021). Fast stochastic bregman gradient methods: Sharp analysis and variance reduction. *ICML*.

Dvurechensky, P., Ostroukhov, P., Safin, K., Shtern, S., and Staudigl, M. (2020). Self-concordant analysis of frank-wolfe algorithms. In *International Conference on Machine Learning*, pages 2814–2824. PMLR.

D'Orazio, R., Loizou, N., Laradji, I., and Mitliagkas, I. (2021). On stochastic mirror descent: Convergence analysis and adaptive variants. *preprint*.

Eckstein, J. (1993). Nonlinear proximal point algorithms using bregman functions, with applications to convex programming. *Mathematics of Operations Research*, 18(1):202–226.

Eckstein, J. (1998). Approximate iterations in bregman-function-based proximal algorithms. *Mathematical programming*, 83(1):113–123.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR.

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

Hanzely, F., Richtarik, P., and Xiao, L. (2021). Accelerated bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79(2):405–440.

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*.

Lu, H. (2019). Relative Continuity for non-lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent. *INFORMS Journal on Optimization*, 1(4):288–303.

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.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, volume 37. CRC Press.

Ostrovskii, D. M. and Bach, F. (2021). Finite-sample analysis of M-estimators using self-concordance. *Electronic Journal of Statistics*, 15(1).

Pav, S. E. (2015). Moments of the log non-central chi-square distribution. *arXiv preprint arXiv:1503.06266*.

Pfau, D. (2013). A generalized bias-variance decomposition for bregman divergences. http://davidpfau.com/assets/generalized_bvd_proof.pdf. [Online; accessed February 23rd 2021].

Raskutti, G. and Mukherjee, S. (2015). The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457.

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

A Other Exponential Families

A.1 Full Gaussian

To model $\mathcal{N}(m, \sigma^2)$, we take the sufficient statistic $T(X) = (X, X^2)$. Then the mean parameter is $\mu = \mathbb{E}[T(X)] = (m, m^2 + \sigma^2) = (\mu_1, \mu_2)$. The log-density is written

$$\log p(X) = -\frac{(X-m)^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2 + \mathrm{cst} \tag{134}$$

$$= \frac{m}{\sigma^2}X - \frac{1}{2\sigma^2}X^2 - \frac{m^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2 \tag{135}$$

$$= \theta_1 X + \theta_2 X^2 - A(\theta)\ . \tag{136}$$

From this formula, we can identify the duality map

$$\theta_1 = \frac{m}{\sigma^2} = \frac{\mu_1}{\mu_2 - \mu_1^2} \tag{137}$$

$$\theta_2 = -\frac{1}{2\sigma^2} = \frac{-1}{2(\mu_2 - \mu_1^2)} < 0 \tag{138}$$

and we can also explicit the log-partition as a function of the natural parameters

$$A(m, \sigma) = \frac{m^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2 \tag{139}$$

$$A(\theta) = \frac{\theta_1^2}{-4\theta_2} - \frac{1}{2}\log(-\theta_2)\ . \tag{140}$$

We can perform a sanity check by deriving A and verifying that its gradient gives $\mu(\theta)$

$$\mu_1 = \frac{\theta_1}{-2\theta_2} \tag{141}$$

$$\mu_2 = \left(\frac{\theta_1}{-2\theta_2}\right)^2 + \frac{1}{-2\theta_2}\ . \tag{142}$$

Now let us recover the entropy

$$A^*(\mu) = \mu_1\theta_1 + \mu_2\theta_2 - A(\theta(\mu)) \tag{143}$$

$$= \frac{\mu_1^2}{\mu_2 - \mu_1^2} - \frac{\mu_2}{2(\mu_2 - \mu_1^2)} - \frac{\mu_1^2}{2(\mu_2 - \mu_1^2)} - \frac{1}{2}\log(\mu_2 - \mu_1^2) \tag{144}$$

$$= \frac{1}{2}\frac{\mu_1^2 - \mu_2}{\mu_2 - \mu_1^2} - \frac{1}{2}\log(\mu_2 - \mu_1^2) \tag{145}$$

$$= -\frac{1}{2}\log(\mu_2 - \mu_1^2) + \mathrm{cst}\ . \tag{146}$$

In this case, the Bregman divergences induced by the log-partition or the entropy are ugly, so we will not focus on them. Note that this entropy is not self-concordant. Taking derivatives in the direction of μ_1, and setting $\mu = (4, 16)$, the self-concordance inequality is not verified.

B Three points lemma

The typical descent lemma of updates involving minimization of a Bregman divergence is called the three point lemma. It corresponds to first order optimality conditions of the update. With its hep, one can prove convergence for mirror descent with relative smoothness, or for Bregman proximal point algorithms.

Lemma B.1. *Let x_+ be a solution to*

$$x_+ \in \operatorname*{argmin}_{x \in C} \overbrace{f(x) + \mathcal{B}_h(x; y)}^{\phi(x)} \tag{147}$$

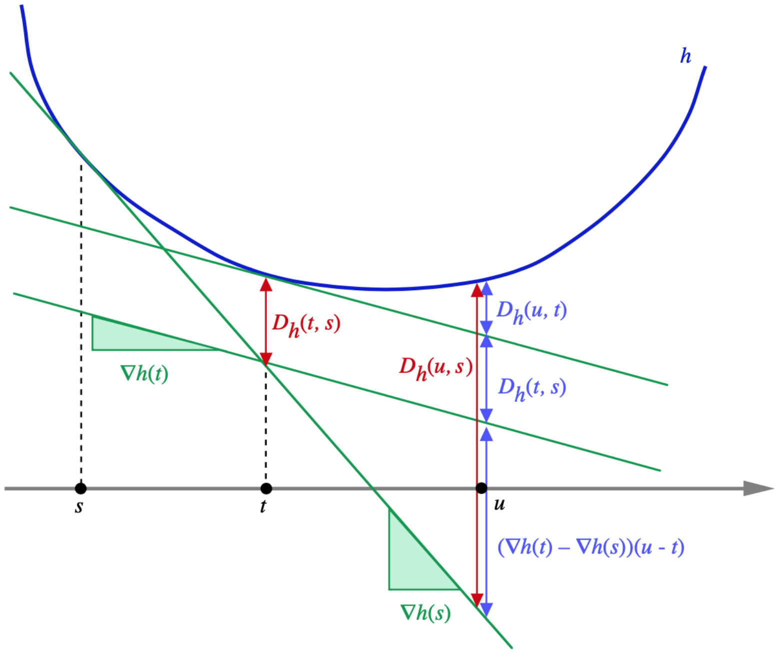

Figure 9: Illustration of the other three points lemma – eg the identity between Bregmans – taken from Eckstein (1998).

where C is a closed convex set, f is a convex function, $\mathcal{B}_h$ is the Bregman divergence induce by some convex function h, and y is some reference vector. Then

$$\forall x, f(x) + \mathcal{B}(x; y) \geq f(x_+) + \mathcal{B}(x; x_+) + \mathcal{B}(x_+; y) \ . \tag{148}$$

Taking an information geometry lens, this property is analog to the Pythagorean theorem for generalized projections. We recover euclidean projections setting $f = 0$ and $h = \|.\|^2$.

Proof. The first order optimality condition is

$$\langle \nabla \phi(x_+), x - x_+ \rangle \geq 0, \forall x \tag{149}$$

$$= \langle \nabla f(x_+) + \nabla h(x_+) - \nabla h(y), x - x_+ \rangle \tag{150}$$

This proof relies on another property called three point property

$$\langle \nabla h(x_+) - \nabla h(y), x - x_+ \rangle = \mathcal{B}(x; y) - \mathcal{B}(x; x_+) - \mathcal{B}(x_+; y) \tag{151}$$

which can be proved by expanding the right hand side. By convexity of f we also have

$$f(x_+) + \langle \nabla f(x_+), x - x_+ \rangle \leq f(x) \ . \tag{152}$$

Putting it all together we get

$$0 \leq \langle \nabla f(x_+), x - x_+ \rangle + \langle \nabla h(x_+) - \nabla h(y), x - x_+ \rangle \tag{153}$$

$$\leq f(x) - f(x_+) + \mathcal{B}(x; y) - \mathcal{B}(x; x_+) - \mathcal{B}(x_+; y) \tag{154}$$

which concludes the proof. □

C Relative Smoothness for Mirror Descent

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.

$$\begin{aligned} & f \text{ is } L\text{-smooth relative to } h && (155)\\ \iff & Lh - f \text{ convex} && (156)\\ \iff & f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + L\mathcal{B}_h(y;x), \forall x, y && (157)\\ \iff & \mathcal{B}_f(y;x) \leq L\mathcal{B}_h(y;x), \forall x, y && (158)\\ \iff & \nabla^2 f(x) \leq L\nabla^2 h(x), \forall x && (159)\end{aligned}$$

where the last equivalence holds only when f and h are twice differentiable. In words, f is upper bounded by it linear approximation plus the h-Bregman divergence, which can also be seen as a bound between divergences, or more locally as a bound between Hessians. Similarly, f is μ strongly-convex relative to h if

$$\begin{aligned} & f - \mu h \;\text{ convex} && (160)\\ \iff & f(x) + \langle \nabla f(x), y - x\rangle + \mu\mathcal{B}_h(y;x) \leq f(y)\forall x, y && (161)\\ \iff & \mu\mathcal{B}_h(y;x) \leq \mathcal{B}_f(y;x)\forall x, y && (162)\\ \iff & \mu\nabla^2 h(x) \leq \nabla^2 f(x)\forall x\;. && (163)\end{aligned}$$

Another way to view this elegant generalization of smoothness and strong-convexity is as a transfer of the Loewner partial order on symmetric 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 that when f is L-smooth relative to h, then the suboptimality of the sequence

$$x_{t+1} = \operatorname*{argmin}_x \langle \nabla f(x_t), x\rangle + \mathcal{B}_h(x; x_t) \tag{164}$$

is upper bounded by the simple formula

$$\implies f(x_t) - f(x_*) \leq \frac{L\mathcal{B}_h(x_*; x_0)}{t}\;. \tag{165}$$

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.

C.1 Convergence of Mirror Descent

Descent Lemma. To get the descent lemma, apply successively relative smoothness, 3 points lemma, and relative strong convexity on f. Note that the strong convexity constant m may be zero. Recall that $\ell_f(\theta, \theta^t) = f(\theta^t) + \langle \nabla f(\theta^t); \theta - \theta^t\rangle$ is the linearization of f at θ^t evaluated in θ.

$$\begin{aligned} f(\theta^{t+1}) &\leq \ell_f(\theta^{t+1}, \theta^t) + L\mathcal{B}_A(\theta^{t+1}; \theta^t) && \text{(Relative } L\text{-smoothness)} \quad (166)\\ &\leq \ell_f(\theta^*, \theta^t) + L\mathcal{B}_A(\theta^*; \theta^t) - L\mathcal{B}_A(\theta^*; \theta^{t+1}) && \text{(3 Points Lemma w. } x = \theta^*) \quad (167)\\ &\leq f(\theta^*) + (L - m)\mathcal{B}_A(\theta^*; \theta^t) - L\mathcal{B}_A(\theta^*; \theta^{t+1}) && \text{(Relative } m\text{-strong convexity)} \quad (168)\end{aligned}$$

Monotonicity. Applying the 3 points lemma with $x = \theta_t$,

$$f(\theta^{t+1}) \leq f(\theta^t) - L(\mathcal{B}_A(\theta^t, \theta^{t+1}) + \mathcal{B}_A(\theta^{t+1}, \theta^t)) \tag{169}$$

shows the function value is monotonically decreasing, and we the gap at each step is greater than the symmetrized bregman between iterates.

No Strong-Convexity. The simpler case is when $m = 0$, then we get a telescopic sum

$$\sum_{t=0}^{T-1} (f(\theta^{t+1}) - f(\theta^*)) \leq L \sum_t \mathcal{B}_A(\theta^*; \theta^t) - \mathcal{B}_A(\theta^*); \theta^{t+1}) \tag{170}$$

$$= L\mathcal{B}_A(\theta^*; \theta^0) - L\mathcal{B}_A(\theta^*); \theta^T) \tag{171}$$

$$\leq L\mathcal{B}_A(\theta^*; \theta^0) , \tag{172}$$

from which we can conclude thanks to the monotonicity of $f(\theta^t)$

$$f(\theta^T) - f(\theta^*) \leq \frac{L}{T}\mathcal{B}_A(\theta^*; \theta^0) , \tag{173}$$

Strong Convexity. When $m > 0$, we need to do some cooking with a coefficient $\frac{L-m}{L}$ to create the telescopic sum

$$\sum_{t=0}^{T-1} \left(\frac{L-m}{L}\right)^{T-1-t} (f(\theta^{t+1}) - f(\theta^*)) \leq \sum_{t=0}^{T-1} \frac{(L-m)^{T-t}}{L^{T-t-1}}\mathcal{B}_A(\theta^*; \theta^t) - \frac{(L-m)^{T-t-1}}{L^{T-t-2}}\mathcal{B}_A(\theta^*); \theta^{t+1}) \tag{174}$$

$$= \frac{(L-m)^T}{L^{T-1}}\mathcal{B}_A(\theta^*; \theta^0) - L\mathcal{B}_A(\theta^*); \theta^T) \tag{175}$$

$$\leq L\left(\frac{L-m}{L}\right)^T \mathcal{B}_A(\theta^*; \theta^0) . \tag{176}$$

Then using the monotonicity of $f(\theta^t)$, we end up with the geometric sum of $\frac{L-m}{L} < 1$

$$\sum_{t=0}^{T-1} \left(\frac{L-m}{L}\right)^{T-1-t} (f(\theta^{t+1}) - f(\theta^*)) \geq (f(\theta^T) - f(\theta^*)) \sum_{t=0}^{T-1} \left(\frac{L-m}{L}\right)^{T-1-t} \tag{177}$$

$$\geq (f(\theta^T) - f(\theta^*)) \frac{1}{1 - \frac{L-m}{L}} \tag{178}$$

$$= (f(\theta^T) - f(\theta^*)) \frac{L}{m} . \tag{179}$$

Finally we end up with the final iterate linear convergence rate

$$f(\theta^T) - f(\theta^*) \leq m\left(1 - \frac{m}{L}\right)^T \mathcal{B}_A(\theta^*; \theta^0) \tag{180}$$

featuring $\frac{L}{m}$ the relative condition number.

D Cramer-Rao Lower Bound

In statistics, there is this famous theorem known as the Cramer-Rao lower bound. In its simplest form, it states that the covariance of an unbiased estimator of a distribution's parameter θ is lower bounded by the inverse Fisher information matrix, divided by n

$$\mathrm{Cov}(\hat{\theta}_n) \succeq \frac{1}{n}\mathcal{I}(\theta_*)^{-1} \tag{181}$$

$$\iff \mathbb{E}_{\mathcal{D}\sim\theta_*}\left[(\hat{\theta}_n - \theta_*)(\hat{\theta}_n - \theta_*)^T\right] \succeq \frac{1}{n}\mathbb{E}_{X\sim\theta_*}\left[-\nabla^2 \log p(X|\theta_*)\right]^{-1} . \tag{182}$$

This is a typical statistic bound, which is concerned with the ℓ^2 accuracy of an estimator, via its covariance. In contrast, here we are concerned with the quality of the estimator in terms of KL-divergence, eg "how accurately am I modeling this distribution?"

Note that in the case of an exponential family, the lower bound is always reached for the MLE estimator $\hat{\mu}_n$, because $\mathcal{I}(\mu) = \nabla^2 A^*(\mu) = \nabla^2 A(\theta)^{-1} = \mathrm{Cov}(T(X))^{-1}$, where the first step takes some calculus. For the estimator $\hat{\theta}_n$, this lower bound is not trivial.

In our case, the bias variance decomposition tells us that we are interested in the covariance between primal and dual parameters, rather than the covariance of only of these – $\mathrm{Cov}(\hat{\theta}_n; \hat{\mu}_n)$ rather that $\mathrm{Var}(\hat{\mu}_n)$ or $\mathrm{Var}(\hat{\theta}_n)$.

E Posterior Expected Loss

A conjugate prior for the exponential family with sufficient statistic $T(X)$ and natural parameter θ is written as

$$p(\theta) \propto \exp(-n_0 \mathcal{B}_A(\theta;\theta_0)) \propto \exp(-n_0 A(\theta) + \langle n_0\mu_0;\theta\rangle) \,. \tag{183}$$

We see that $p(\theta)$ is part of the exponential family, with sufficient statistic $(\theta, A(\theta))$ and natural parameters $(n_0\mu_0, -n_0)$. When we update this prior with Bayes formula on dataset $\mathcal{D} = (x_1, \dots, x_n)$, we get a posterior $p(\theta|\mathcal{D})$ with natural parameters $(n_0\mu_0 + \sum_i T(x_i), -(n_0+n))$. This is well detailed in Agarwal and Daumé (2010).

Given this posterior, a natural quantity to consider is the expected loss. Thus we might ask "what is the expectation of the posterior expected suboptimality?"

$$\mathbb{E}_{\mathcal{D}\sim\theta_*}\left[\mathbb{E}_{p(\theta|\mathcal{D})}\left[\mathcal{B}_A(\theta;\theta_*)\right]\right] \tag{184}$$

This quantity is even harder to estimate than the expected MAP suboptimality, because of the double integral. We can get a closed form for the age-old ℓ^2 loss of $\mathcal{N}(\mu,1)$, which yields the exact rate $\frac{1}{n}$. This is exactly twice the expected loss of the MLE.

We can also get a closed form for the posterior of a centered gaussian variance

$$\mathbb{E}_{p(\theta|\mathcal{D})}\left[\mathcal{B}_A(\theta;\theta_*)\right] = \mathcal{B}_A(\hat{\theta}_n;\theta_*) + \frac{2}{n}\frac{\mu_*}{\hat{\mu}_n} + \log\frac{n}{2} - \psi(1+\frac{n}{2}) \tag{185}$$

using the mean of the inverse Gamma distribution and the logarithmic expectation of a Gamma. Then using results from the MLE, we get the exact formula

$$\mathbb{E}_{\mathcal{D}\sim\theta_*}\left[\mathbb{E}_{p(\theta|\mathcal{D})}\left[\mathcal{B}_A(\theta;\theta_*)\right]\right] = \frac{2}{n} + \frac{2}{n(n-2)} + \psi(\frac{n}{2}) - \psi(1+\frac{n}{2}) \tag{186}$$

and we bound it to

$$\frac{1}{2n} + \frac{4}{n(n-2)} \leq \mathbb{E}_{\mathcal{D}\sim\theta_*}\left[\mathbb{E}_{p(\theta|\mathcal{D})}\left[\mathcal{B}_A(\theta;\theta_*)\right]\right] \leq \frac{1}{n} + \frac{4}{n(n-2)} \tag{187}$$

which is about twice the MLE bounds (35), exactly as we reportred for the gaussian mean.

F Gaussian Variance

F.1 Proof of MLE Tight Bound

Theorem F.1 (MLE tight upper bound). *The MLE of $\mathcal{N}(0,\mu_*)$ is $\mu_n = \frac{1}{n}\sum_i X_i^2$. Its expected suboptimality is infinite when $n \leq 2$ and otherwise*

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*;\mu_n)\right] \leq \frac{1}{2n} + \frac{2}{n(n-2)} \,. \tag{188}$$

Proof. The ratio to the optimum $\frac{\mu_n}{\mu_*}$follows a Chi-square distribution with n degrees of freedom $\chi^2(n)$ divided by n. Its inverse $\frac{\mu_*}{\mu_n}$ follows an inverse Chi-square distribution with expectation

$$\mathbb{E}\left[\frac{\mu_*}{\mu_n} - 1\right] = \mathbb{E}\left[\frac{n}{\chi^2(n)} - 1\right] = \begin{cases} \frac{n}{n-2} - 1 = \frac{2}{n-2} & \text{if } n > 2, \\ +\infty & \text{otherwise.} \end{cases} \tag{189}$$

There is also a closed form solution for the expected logarithm of a Chi-squared(Pav, 2015)

$$\mathbb{E}\left[\log\frac{\mu_n}{\mu_*}\right] = \psi(\frac{n}{2}) - \log(\frac{n}{2}) \tag{190}$$

where ψ is the digamma function. Consequently the suboptimality of the MLE has a closed form solution

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*;\mu_n)\right] = \frac{1}{2}\mathbb{E}\left[\frac{\mu_*}{\mu_n} - 1 + \log\left(\frac{\mu_n}{\mu_*}\right)\right] \tag{191}$$

$$= \begin{cases} \frac{1}{2}\left(\frac{2}{n-2} + \psi(\frac{n}{2}) - \log(\frac{n}{2})\right) & \text{if } n > 2, \\ +\infty \quad \text{otherwise.} \end{cases} \tag{192}$$

It is surprising that we need more than 3 samples for the loss to have a bounded expectation. It is also very pleasant that the optimal value μ_* does not appear in this rate, so that the convergence rate is independent of the actual solution. When the expectation is finite, we can make its formula intelligible thanks to bounds on the digamma function

$$-\frac{1}{x} \leq \psi(x) - \log(x) \leq -\frac{1}{2x} \; . \tag{193}$$

For $n \geq 3$,

$$\frac{1}{n-2} - \frac{1}{n} \leq \mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*;\mu_n)\right] \leq \frac{1}{n-2} - \frac{1}{2n} \tag{194}$$

$$\iff \frac{2}{n(n-2)} \leq \mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*;\mu_n)\right] \leq \frac{1}{2n} + \frac{2}{n(n-2)} \tag{195}$$

so we get a $O(n^{-2})$ lower bound and a $O(n^{-1}) + O(n^{-2})$ upper bound. □

Using the Stirling series, we can also get an asymptotic approximation

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*;\mu_n)\right] \in \frac{1}{2n} + \frac{11}{6n^2} + \frac{4}{n^2(n-2)} + O(n^{-4}) \; . \tag{196}$$

F.2 Proof of MLE Loose Upper Bound

If we apply (34) to the MLE, we get the upper bound

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*;\hat{\mu}_n^{\text{MLE}})\right] \leq \frac{1}{2}\left(\mathbb{E}\left[\frac{n}{\chi^2(n)}\right] + \mathbb{E}\left[\frac{\chi^2(n)}{n}\right] - 2\right) \tag{197}$$

$$= \frac{1}{2}\left(\frac{n}{n-2} + 1 - 2\right) = \frac{1}{n-2} \tag{198}$$

which is exactly an additive term $1/2n$ larger than (35). This highlights the factor 2 difference between the log-bound and an exact computation of the expectation of the log term, and a tight upper bound on the digamma function.

F.3 Multivariate MLE

In higher dimensions d

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*;\hat{\mu}_n^{\text{MLE(d)}})\right] = \frac{1}{2}\left(\frac{d(d+1)}{n-d-1} + \psi_d(\frac{n}{2}) - d\log(\frac{n}{2})\right) \tag{199}$$

where $\psi_d(\frac{n}{2}) = \sum_{i=0}^{d-1}\psi(\frac{n-i}{2})$ is the multivariate digamma function. Using the same upper bound as in the univariate case

$$\psi_d(\frac{n}{2}) - d\log(\frac{n}{2}) \leq \sum_{i=0}^{d-1}\log(1-\frac{i}{n}) - \frac{1}{n-i} \; . \tag{200}$$

We can bound the harmonic sum with typical bounds

$$-\sum_{i=0}^{d-1}\frac{1}{n-i} = H_{n-d} - H_n \tag{201}$$

$$\leq \log(n-d) + \gamma + \frac{1}{2(n-d)-1} - \log(n) - \gamma - \frac{1}{2n+1} \tag{202}$$

$$= \log(1-\frac{d}{n}) + \frac{2(d+1)}{(2n+1)(2(n-d)-1} \tag{203}$$

which yields

$$\psi_d(\frac{n}{2}) - d\log(\frac{n}{2}) \leq \sum_{i=0}^{d}\log(1-\frac{i}{n}) + \frac{2(d+1)}{(2n+1)(2(n-d)-1}\,. \tag{204}$$

From there we have two options. Either we use a concave tangent bound on the logarithms to get

$$\sum_{i=0}^{d}\log(1-\frac{i}{n}) \leq \sum_{i=0}^{d} -\frac{i}{n} = -\frac{d(d+1)}{2n} \tag{205}$$

which yields the final rate

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*;\hat{\mu}_n^{\mathrm{MLE(d)}})\right] \leq \frac{d(d+1)}{4n} + \frac{d(d+1)^2+\frac{d+1}{2}}{2n(n-d-1)}\,. \tag{206}$$

Note that this rate is not tight for $d>1$, but it is tight for $d=1$, and we almost recover the previous rate.

Another option, potentially tighter than the tangent bound, is to bound the sum of logarithms with an integral, using the monotonicity of the logarithm $\log(1-\frac{i}{n}) \leq \log(1-\frac{x}{n}), \forall x \in [i-1,i]$

$$\sum_{i=1}^{d}\log(1-\frac{i}{n}) \leq \int_0^d \log(1-\frac{x}{n})dx \tag{207}$$

$$= n\int_{1-\frac{d}{n}}^{1}\log(y)dy \tag{208}$$

$$= n[y(\log(y)-1)]_{1-\frac{d}{n}}^{1} \tag{209}$$

$$= -n-(n-d)(\log(1-\frac{d}{n})-1) \tag{210}$$

$$= -(n-d)\log(1-\frac{d}{n}) - d\,. \tag{211}$$

Unfortunately, it is hard to get a tractable and non-vacuous bound for this guy. At $n=d$ it is worth $-d$, and then it grows up to 0. and we can bound it using some exponents of $1-\frac{d}{n}$, but that is not really helpful to get an easy formula. Actually looking at graphics, the first upper bound turns out to be tighter than the second one.

F.4 Proof of Expected MAP Natural Parameter

Let us introduce a variable characterizing the importance of the prior

$$a = n_0\frac{\mu_0}{\mu^*}\,. \tag{212}$$

Lemma F.2 (Expected MAP natural parameter)**.** *The expectation of the natural parameter of the MAP of $\mathcal{N}(0,\mu_*)$ is bounded as*

$$\frac{\mu^*}{\mu_n} \leq \mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] = \mathbb{E}\left[\frac{\hat{\theta}_n}{\theta^*}\right] \leq \begin{cases}\frac{n_0+n}{a+n-2} & \textit{when } n\geq 2,\\ \frac{n_0+1}{a} & \textit{when } n=1.\end{cases} \tag{213}$$

Proof. The lower bound can be readily obtained with the Jensen inequality applied on the convex function $x \mapsto \frac{1}{x}$ for $x > 0$. The upper bound requires much more work. To start, let us plug in the definition of $\hat{\mu}_n$

$$\mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] = \mathbb{E}\left[\frac{\hat{\theta}_n}{\theta^*}\right] = \mathbb{E}\left[\frac{(n_0+n)\mu^*}{n_0\mu_0 + \sum_i X_i^2}\right] \tag{214}$$

$$= (n_0+n)\mathbb{E}\left[\frac{1}{n_0\frac{\mu_0}{\mu^*} + \sum_i \frac{X_i^2}{\mu^*}}\right] \tag{215}$$

$$= (n_0+n)\mathbb{E}\left[\frac{1}{a+\chi^2(n)}\right] \tag{216}$$

where $\chi^2(n) = \sum_i \frac{X_i^2}{\mu^*}$ is a chi-square random variable of degree n and $a = n_0\frac{\mu_0}{\mu^*}$. We reformulate this expectation with the trick

$$\int_1^\infty e^{-\frac{zt}{2}}\,dt = \frac{2}{z}e^{-\frac{z}{2}} \tag{217}$$

$$\implies \frac{1}{z} = \frac{1}{2}e^{\frac{z}{2}}\int_1^\infty e^{-\frac{zt}{2}}\,dt \tag{218}$$

$$\implies \frac{1}{a+x} = \frac{1}{2}e^{\frac{a+x}{2}}\int_1^\infty e^{-\frac{(a+x)t}{2}}\,dt\ , \tag{219}$$

to get two integrals instead of one

$$2^{\frac{n}{2}}\Gamma(\frac{n}{2})\mathbb{E}\left[\frac{1}{a+\chi^2(n)}\right] = \int_0^\infty \frac{x^{\frac{n}{2}-1}e^{-\frac{x}{2}}}{a+x}dx \tag{220}$$

$$= \int_0^\infty dx\ x^{\frac{n}{2}-1}e^{-\frac{x}{2}}\frac{1}{2}e^{\frac{a+x}{2}}\int_1^\infty dt\ e^{-\frac{(a+x)t}{2}} \tag{221}$$

$$= \frac{1}{2}e^{\frac{a}{2}}\int_1^\infty dt\ e^{-\frac{at}{2}}\int_0^\infty dx\ x^{\frac{n}{2}-1}e^{-\frac{xt}{2}}\ , \tag{222}$$

where we used Fubini to switch integrals, without further justification. Now with the change of variable $x = 2\frac{y}{t}$, eg $dx = 2\frac{dy}{t}$, the inner integral becomes

$$\mathbb{E}\left[\frac{1}{a+\chi^2(n)}\right] = \frac{e^{\frac{a}{2}}}{2^{\frac{n}{2}+1}\Gamma(\frac{n}{2})}\int_1^\infty dt\ e^{-\frac{at}{2}}\int_0^\infty 2\frac{dy}{t}\ (2\frac{y}{t})^{\frac{n}{2}-1}e^{-y} \tag{223}$$

$$= \frac{e^{\frac{a}{2}}}{2\Gamma(\frac{n}{2})}\underbrace{\int_1^\infty \frac{e^{-\frac{at}{2}}}{t^{\frac{n}{2}}}dt}_{E_{\frac{n}{2}}(\frac{a}{2})}\ \underbrace{\int_0^\infty y^{\frac{n}{2}-1}e^{-y}dy}_{\Gamma(\frac{n}{2})} \tag{224}$$

so we finally get the formula valid for all $n \geq 1$

$$\boxed{\mathbb{E}\left[\frac{1}{a+\chi^2(n)}\right] = \frac{1}{2}e^{\frac{a}{2}}E_{\frac{n}{2}}(\frac{a}{2})} \tag{225}$$

where the generalized exponential integral function is defined as

$$E_k(z) = \int_1^\infty \frac{e^{-zt}}{t^k}dt\ . \tag{226}$$

Now our goal is to bound this function with simpler functions. Fortunately, mathematicians have been working on these integrals for decades. When $n = 2$, eg $k = 1$, we have the simple exponential integral function which verifies the bound

$$\frac{1}{2}\log(1+\frac{2}{x}) \leq e^x E_1(x) \leq \log(1+\frac{1}{x}) \leq \frac{1}{x}\ . \tag{227}$$

For $n \geq 2$, eg $k \geq 1$, we have the general bound

$$\frac{1}{x+k} \leq e^x E_k(x) \leq \frac{1}{x+k-1} \tag{228}$$

so that

$$\begin{aligned}
\frac{1}{a+n} &\leq \mathbb{E}\left[\frac{1}{a+\chi^2(n)}\right] \leq \frac{1}{a+n-2} && (229)\\
\iff \frac{n+n_0}{a+n} &\leq \mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] \leq \frac{n+n_0}{a+n-2} && (230)\\
\iff \frac{\mu^*}{\mu_n} &\leq \mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] \leq \frac{1}{\frac{\mu_n}{\mu^*} - \frac{2}{n+n_0}} && (231)
\end{aligned}$$

where we used $\frac{a+n}{n+n_0} = \frac{\mu_n}{\mu^*}$.

We are left with a special case when $n = 1$, eg $k = \frac{1}{2}$. Then we can use the trivial bound

$$\mathbb{E}\left[\frac{1}{a+X^2}\right] < \frac{1}{a} \tag{232}$$

which is what we are using in the lemma, or we can get a stronger result with further derivations. The integral can be written with Mill's ratio $M(x) = e^{x^2} \int_x^\infty e^{-t^2} dt$, by doing the change of variable $t = x\sqrt{u}$, $dt = \frac{x}{2}\frac{du}{\sqrt{u}}$

$$\begin{aligned}
\int_x^\infty e^{-t^2} dt &= \frac{x}{2}\int_1^\infty \frac{e^{-x^2 u}}{\sqrt{u}} du = \frac{x}{2} E_{\frac{1}{2}}(x^2) && (233)\\
\implies M(x) &= \frac{x}{2} e^{x^2} E_{\frac{1}{2}}(x^2) && (234)\\
\implies e^y E_{\frac{1}{2}}(y) &= 2\frac{M(\sqrt{y})}{\sqrt{y}} && (235)
\end{aligned}$$

so that

$$\mathbb{E}\left[\frac{1}{a+X^2}\right] = \frac{1}{2} e^{\frac{a}{2}} E_{\frac{1}{2}}(\frac{a}{2}) = \sqrt{\frac{2}{a}} M(\sqrt{\frac{a}{2}}) = \frac{2}{a}\sqrt{\frac{a}{2}} M(\sqrt{\frac{a}{2}}) \,. \tag{236}$$

Then using another bound from the DLMF, we get that

$$\frac{1}{a+1} \leq \mathbb{E}\left[\frac{1}{a+X^2}\right] < \frac{1}{a}\left(1 - \frac{1}{a+3}\right) \,. \tag{237}$$

□

F.5 Proof of MAP Bound

Theorem F.3 (MAP Upper Bound). *The expected suboptimality of the MAP of $\mathcal{N}(0, \mu^*)$ with prior hyper-parameters (n_0, μ_0) is*

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu_*; \hat{\mu}_n^{MAP})\right] \leq \begin{cases} \mathcal{B}_{A^*}(\mu_*; \mu_0) \quad \textit{if} \quad n = 0, \\ \frac{1}{2(n_0+1)} + \frac{(1+\frac{1}{n_0} - \frac{\mu_0}{\mu^*})^2}{2\frac{\mu_0}{\mu^*}(1+\frac{1}{n_0})} \;\; \textit{if } n = 1, \\ \frac{1}{n_0\frac{\mu_0}{\mu^*}+n-2} + \frac{(1+\frac{1}{n_0} - \frac{\mu_0}{\mu^*})^2}{2(\frac{\mu_0}{\mu^*}+\frac{n-2}{n_0})(1+\frac{n}{n_0})} \;\; \textit{if } n \geq 2 \end{cases} \tag{238}$$

Proof. When $n = 0$, the inequality is an equality. Wen $n > 0$, Wolfram finds no closed form for $\mathbb{E}\left[\log \frac{\hat{\mu}_n}{\mu^*}\right]$ so we focus on the simple log-bound given by (34),

$$2\mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*; \hat{\mu}_n)\right] \leq \mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] - 1 + \mathbb{E}\left[\frac{\hat{\mu}_n}{\mu^*}\right] - 1 \,. \tag{239}$$

We readily get

$$\mathbb{E}\left[\frac{\hat{\mu}_n}{\mu^*}\right] - 1 = \frac{a+n}{n_0+n} - 1 = \frac{a-n_0}{n_0+n} \tag{240}$$

where recall that $a = n_0\frac{\mu_0}{\mu^*}$. There remains the more problematic term with the expectation of the inverse mean parameter. Fortunately, we derived the bound (213) for this purpose.

When $n \geq 2$, we get

$$\mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] - 1 \leq \frac{n_0+n}{a+n-2} - 1 = \frac{n_0-a+2}{a+n-2} = \frac{2}{a+n-2} + \frac{n_0-a}{a+n-2} \tag{241}$$

so putting it all together we get

$$\begin{aligned}
\mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*;\hat{\mu}_n)\right] &\leq \frac{1}{a+n-2} + \frac{n_0-a}{2}\left(\frac{1}{a+n-2} - \frac{1}{n_0+n}\right) && (242)\\
&\leq \frac{1}{a+n-2} + \frac{n_0-a}{2}\,\frac{n_0-a+2}{(a+n-2)(n_0+n)} && (243)\\
&= \frac{1}{a+n-2} + \frac{(n_0-a)^2+2(n_0-a)\pm 1}{2(a+n-2)(n_0+n)} && (244)\\
&= \frac{1}{a+n-2} + \frac{(n_0-a+1)^2-1}{2(a+n-2)(n_0+n)} && (245)\\
&\leq \frac{1}{a+n-2} + \frac{(n_0-a+1)^2}{2(a+n-2)(n_0+n)} && (246)\\
&= \frac{1}{a+n-2} + \frac{(1+\frac{1}{n_0}-\frac{\mu_0}{\mu^*})^2}{2(\frac{\mu_0}{\mu^*}+\frac{n-2}{n_0})(1+\frac{n}{n_0})}\ . && (247)
\end{aligned}$$

When $n = 1$ the bound (213) on the expected natural parameter gives

$$\mathbb{E}\left[\frac{\mu^*}{\hat{\mu}_n}\right] - 1 \leq \frac{n_0+1}{a} - 1 = \frac{n_0+1-a}{a} \tag{248}$$

so putting it all together we get

$$\begin{aligned}
2\mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*;\hat{\mu}_1)\right] &\leq \frac{a-n_0\pm 1}{n_0+1} + \frac{n_0-a+1}{a} && (249)\\
&= \frac{1}{n_0+1} + (a-n_0-1)(\frac{1}{n_0+1} - \frac{1}{a}) && (250)\\
&= \frac{1}{n_0+1} + \frac{(n_0+1-a)^2}{a(n_0+1)} && (251)\\
&= \frac{1}{n_0+1} + \frac{(1+\frac{1}{n_0}-\frac{\mu_0}{\mu^*})^2}{2\frac{\mu_0}{\mu^*}(1+\frac{1}{n_0})} && (252)
\end{aligned}$$

so we do recover the same bias variance structure as when $n \geq 2$. □

G More on the Bias-Variance Decomposition

G.1 Bias Term

The pure bias term measures the divergence between the optimum and the expected MAP estimate $\mathcal{B}_{A^*}(\mu^*;\mu_n)$.

Asymptote. Using the quadratic approximation of the Bregman divergence near μ^*, we get a $O(n^{-2})$ rate for the pure bias term

$$\mathcal{B}_{A^*}(\mu^*;\mu_n) = \frac{\|\mu^*-\mu_0\|^2_{\boldsymbol{\Sigma}^{-1}}}{2(1+\frac{n}{n_0})^{-2}} + O(n^{-3})\ . \tag{253}$$

We can use the classical lower bound on the logarithm to upper bound the pure bias term

$$\begin{aligned}
\mathcal{B}_{A^*}(\mu^*;\mu_n) &= \frac{\mu^*}{\mu_n} - 1 - \log\frac{\mu^*}{\mu_n} && (254)\\
&\leq \frac{(\mu^*-\mu_n)^2}{2\mu^*\mu_n} = \frac{1}{2}(\frac{\mu^*}{\mu_n} + \frac{\mu_n}{\mu^*}) - 1 && (255)\\
&= \frac{1}{2}(\frac{n_0+n}{a+n} - 1 + \frac{a+n}{n_0+n} - 1) && (256)\\
&= \frac{1}{2}(\frac{n_0-a}{a+n} + \frac{a-n_0}{n_0+n}) && (257)\\
&= \frac{n_0-a}{2}(\frac{1}{a+n} - \frac{1}{n_0+n}) && (258)\\
&= \frac{n_0-a}{2}\frac{n_0+n-a-n}{(a+n)(n_0+n)} && (259)\\
&= \frac{(n_0-a)^2}{2(a+n)(n_0+n)} && (260)
\end{aligned}$$

which is a $O(n^{-2})$ term, worth 0 when $n_0 = a$, eg $\mu_0 = \mu^*$, exactly as we expected.

G.2 Variance Term

Asymptote. Using a linear approximation of the mirror map, we get a $O(n^{-1})$ rate for the covariance

$$\begin{aligned}
&\nabla A^*(\mu) = \nabla A^*(\mu^*) + \nabla^2 A^*(\mu^*)(\mu-\mu^*) + O(\|\mu-\mu^*\|^2) && (261)\\
\implies &\hat\theta_n = \theta^* + \Sigma^{-1}(\hat\mu_n - \mu^*) + O(\mathbb{E}[\|\hat\mu_n-\mu^*\|^2]) && (262)\\
\implies &\operatorname{Cov}(\hat\theta_n;T(X)) = \frac{1}{n+n_0}\mathbb{E}[\|T(X)-\mu^*\|^2_{\Sigma^{-1}}] + O(n^{-2}) && (263)\\
\implies &\operatorname{Cov}(\hat\theta_n;T(X)) = \frac{d}{n+n_0} + O(n^{-2})\ . && (264)
\end{aligned}$$

Note that we lost a factor $\frac{1}{2}$ compared to the straightforward asymptotic analysis (53), probably because of the Jensen upper bound.

For the gaussian variance MAP, the variance term is

$$\operatorname{Cov}(\hat\theta_n;\hat\mu_n) = \mathbb{E}[\hat\theta_n\hat\mu_n] - \mathbb{E}[\hat\theta_n]\,\mathbb{E}[\hat\mu_n] = -\frac{1}{2} - \mathbb{E}[\hat\theta_n]\mu_n \qquad (265)$$

where we used the fact that $\theta\mu = -\frac{1}{2}$ for all valid pairs (θ,μ). Plugging-in (36) yield

$$\begin{aligned}
-\frac{1}{2} + \frac{\mu_n}{2\mu_n} \leq \operatorname{Cov}(\hat\theta_n;\hat\mu_n) &\leq -\frac{1}{2} + \frac{1}{2}\frac{a+n}{a+n-2} && (266)\\
\iff 0 \leq \operatorname{Cov}(\hat\theta_n;\hat\mu_n) &\leq \frac{1}{2}\frac{a+n-(a+n-2)}{a+n-2} && (267)\\
\iff 0 \leq \operatorname{Cov}(\hat\theta_n;\hat\mu_n) &\leq \frac{1}{a+n-2} && (268)
\end{aligned}$$

which is exactly the kind of upper bounds we were looking for !

G.3 Mixed Bias-Variance Term

The mixed bias variance term is $\frac{n_0}{n+n_0}\langle\mu_* - \mu_0;\theta_n - \mathbb{E}[\hat\theta_n]\rangle$. Note that it might be negative ! Let's focus on the right factor

$$\theta_n - \mathbb{E}\left[\hat\theta_n\right] = \nabla A^*(\mathbb{E}[\hat\mu_n]) - \mathbb{E}[\nabla A^*(\hat\mu_n)] \qquad (269)$$

is sort of a commuting bracket between ∇A^* and the expectation over $\hat{\mu}_n$. It will be large if ∇A^* is highly non linear over the distribution of $\hat{\mu}_n$.

Asymptote. A linear approximation of ∇A^* shows that the mixed term is in $O(n^{-3})$, since the linear function commutes with the expectation. We need to use a quadratic approximation of ∇A^* to get a non-trivial approximation

$$\nabla A^*(\mu) = \nabla A^*(\mu^*) + \nabla^2 A^*(\mu^*)(\mu - \mu^*) + \frac{1}{2}\nabla^3 A^*(\mu^*)[(\mu - \mu^*); (\mu - \mu^*)] + O(\|\mu - \mu^*\|^3) \quad (270)$$

$$\implies \nabla A^*(\mathbb{E}[\hat{\mu}_n]) - \mathbb{E}[\nabla A^*(\hat{\mu}_n)] = \frac{1}{2}\nabla^3 A^*(\mu^*) \operatorname{Cov}(\hat{\mu}_n) + O(\mathbb{E}[\|\hat{\mu}_n - \mu^*\|^3]) \quad (271)$$

$$\implies \nabla A^*(\mathbb{E}[\hat{\mu}_n]) - \mathbb{E}[\nabla A^*(\hat{\mu}_n)] = \frac{\nabla^3 A^*(\mu^*)\Sigma}{2(n+n_0)^2} + O(n^{-3}) \quad (272)$$

$$(273)$$

where the axis of the tensor product between third derivative and vector / matrix are all kept implicit for simplicity. Consequently the bias-variance term is

$$\frac{n_0}{n+n_0}\langle \underbrace{\mu^* - \mu_0}_{\text{bias}}; \underbrace{\theta_n - \mathbb{E}[\hat{\theta}_n]}_{\text{variance}}\rangle = \frac{n_0}{(n+n_0)^3}(\mu^* - \mu_0)^\top \nabla^3 A^*(\mu^*)\Sigma + O(n^{-4}) = O(n^{-3}) \,. \quad (274)$$

As we have the tight bounds (36) for the expectation of the natural parameter of the MAP, we can bound above and below this mixed term

$$0 \le \theta_n - \mathbb{E}[\hat{\theta}_n] \le \frac{n_0+n}{2\mu^*}\left(\frac{1}{a+n-2} - \frac{1}{a+n}\right) = \frac{n_0+n}{\mu^*(a+n)(a+n-2)} \in O(n^{-1}) \,. \quad (275)$$

Remark how this asymptotic class differs from the $O(n^{-2})$ I expected. Is this bound less tight than I thought, or are my calculus wrong ? Also remark that θ_n is always greater than $\mathbb{E}[\hat{\theta}_n]$, which is a good sanity-check because $\nabla A^*(\mu) = -\frac{1}{2\mu}$ is a concave function.

The rate on the mixed term depends on the sign of $\mu^* - \mu_0$

$$\frac{n_0}{n+n_0}\langle \mu^* - \mu_0; \theta_n - \mathbb{E}[\hat{\theta}_n]\rangle \le \begin{cases} 0 & \text{if} \quad \mu^* \le \mu_0, \\ \frac{n_0 - a}{(a+n)(a+n-2)} & \text{otherwise.} \end{cases} \quad (276)$$

G.4 Putting it all together

$$\mathbb{E}\left[\mathcal{B}_{A^*}(\mu^*; \hat{\mu}_n)\right] \le \underbrace{\frac{(n_0 - a)^2}{2(a+n)(n_0+n)}}_{\text{bias}} + \frac{\max(n_0 - a, 0)}{(a+n)(a+n-2)} + \underbrace{\frac{1}{a+n-2}}_{\text{variance}} \quad (277)$$

G.5 More Illustrations of the Bias-Variance Decomposition

See Figure 10.

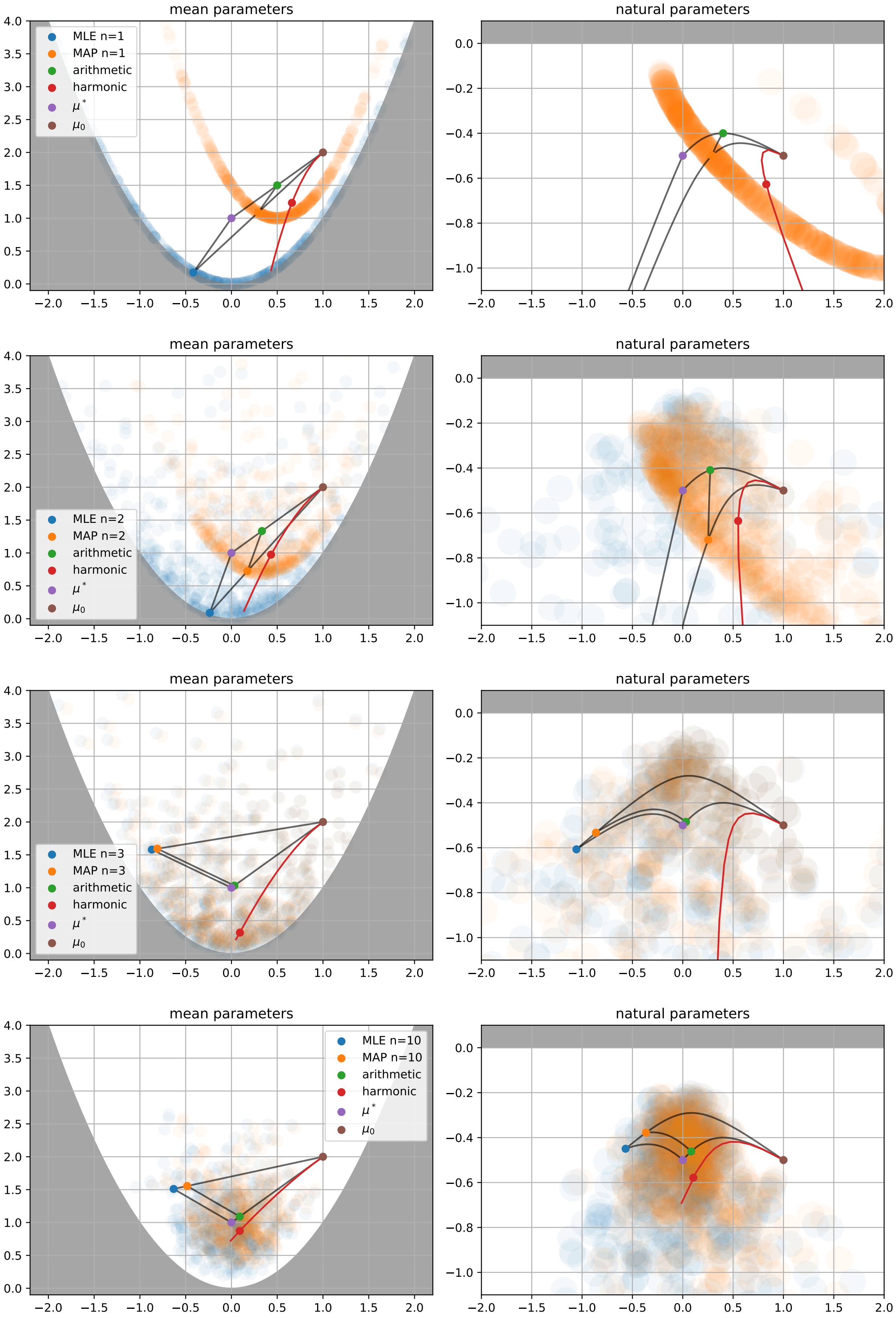

Figure 10: A numerical illustration of the different characters featured in the bias-variance decomposition, for a 1D Gaussian $\mathcal{N}(\mu, \sigma^2)$.

H Self-Concordance

An hypothesis that may be more suitable than smoothness of Lipschitzness is self-concordance. $f : \mathbb{R} \to \mathbb{R}$ is self-concordant (SC) if

$$|f'''(x)| \leq 2f''(x)^{\frac{3}{2}} \ . \tag{278}$$

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. Most importantly, the negative logarithm $-\log(x)$ and its matrix generalization $-\log\det(\boldsymbol{X})$ (to verify) are self-concordant functions. This is great, but it is missing all the other barrier objectives $x^{-\alpha}, \alpha > 0$. For instance, gaussians have a logarithmic term, but hey also have an inverse term which is not self-concordant. These other barriers can fit in the framework of generalized self-concordance (Dvurechensky et al., 2020) , but the calculus quickly gets very complex, as this generalized self-concordance is not as clean as the standard self-concordance.

Questions. Is there already an SC analysis of proximal point methods. (that's not affine invariant)? If the log-partition is SC, then what does it mean for its gradient, the mapping from natural to mean parameters ?

H.1 Damien's Trials

We want to have a rate of convergence on

$$\mathbb{E}_{X_i}\left[\mathcal{B}_{A^*}(\mu_*, \mu_n)\right],$$

where

$$\mu_* = \mathbb{E}_X T(X), \quad \mu_n = \frac{\sum_{i=1}^n T(X_i) + n_0\mu_0}{n_0 + n}.$$

We first assume that A^* is self-concordant.

Notations

- $\| \cdot \|_x = \sqrt{\langle \nabla^2 A^*(x) \cdot, \cdot \rangle}$
- $\| \cdot \|_x^* = \sqrt{\langle [\nabla^2 A^*(x)]^{-1} \cdot, \cdot \rangle}$
- $\lambda(x) = \|(A^*)'(x)\|_x^*$ (Newton decrements)
- $\omega = t - \ln(1+t)$
- $\omega^* = -t - \ln(1-t)$

Note that ω^* is convex and monotonne.

Proposition 1. *(Conversion of norms) We have*

$$\frac{\|y-x\|_x}{1+\|y-x\|_x} \leq \|y-x\|_y \leq \frac{\|y-x\|_x}{1-\|y-x\|_x}$$

Proposition 2. *(Bounded Hessian change) We have*

$$(1-\|y-x\|_x)^2 \nabla^2 A^*(x) \leq \nabla^2 A^*(y) \leq \frac{1}{(1-\|y-x\|_x)^2} \nabla^2 A^*(x)$$

Proposition 3. *(Function bound) We have, if $\|y-x\|_x < 1$,*

$$A^*(y) \leq A^*(x) + \nabla A^*(x)(y-x) + \omega^*(\|y-x\|_x)$$

The last proposition implies that, when $\|y-x\|_x < 1$,

$$\mathcal{B}_{A^*}(y,x) \le \omega^*(\|y-x\|_x).$$

Therefore,

$$\mathcal{B}_{A^*}(\mu^*,\mu_n) \le \omega^*(\|\mu^*-\mu_n\|_{\mu_n}).$$

then, if $\|\mu^*-\mu_n\|_{\mu^*} < 1$, we have

$$\mathcal{B}_{A^*}(\mu^*,\mu_n) \le \omega^*\left(\frac{\|\mu^*-\mu_n\|_{\mu^*}}{1-\|\mu^*-\mu_n\|_{\mu^*}}\right).$$

Assume $\|\mu^*-\mu_n\|_{\mu^*} < c$,

$$\mathcal{B}_{A^*}(\mu^*,\mu_n) \le \omega^*\left(\frac{\|\mu^*-\mu_n\|_{\mu^*}}{1-c}\right).$$

Discarded However,

$$\begin{aligned}\|\mu^*-\mu_n\|^2_{\mu_n} &= \left\|\frac{n+n_0-1}{n+n_0}(\mu^*-\mu_{n-1}) + \frac{1}{n+n_0}(\mu^*-T_n)\right\|^2_{\mu_n},\\ &= \left\|\frac{n+n_0-1}{n+n_0}(\mu^*-\mu_{n-1})\right\|^2_{\mu_n} + \left\|\frac{1}{n+n_0}(\mu^*-T_n)\right\|^2_{\mu_n} + \frac{2(n+n_0+1)}{(n+n_0)^2}\langle\nabla^2[A(\mu_n)](\mu^*-\mu_{n-1}),\,(\mu^*-T_n)\rangle\end{aligned}$$

and

$$\|\mu^*-\mu_n\|_{\mu_n} \le \frac{n+n_0-1}{n+n_0}\left\|\mu^*-\mu_{n-1}\right\|_{\mu_n} + \frac{1}{n+n_0}\left\|\mu^*-T_n\right\|_{\mu_n}.$$

Moreover, by the bounded Hessian change,

$$\|\mu^*-\mu_{n-1}\|_{\mu_n} \le \frac{1}{1-\|\mu_n-\mu_{n-1}\|_{\mu_{n-1}}}\|\mu^*-\mu_{n-1}\|_{\mu_{n-1}}$$