Homeomorphic-Invariance of EM: Non-Asymptotic Convergence in KL Divergence for Exponential Families via Mirror Descent Supplementary Materials

Organization of the supplementary material

Table 1: Summary of notation and acronyms

Context	Symbol	
Data	x, z	Observed (x) and missing (z), or latent, variables.
Parameters	$\theta, \phi \in \Omega$	(Natural) Parameters of the model and set of valid parameters.
	μ	Equivalent mean parameters.
EM	$\mathcal{L}(\theta)$	Objective function, the negative log-likelihood $-\log p(x \mid \theta)$.
	$Q_\theta(\phi)$	Surrogate objective optimized by the M-step.
Exponential families	$S(x, z)$	Sufficient statistics.
	$A(\theta)$, $A^*(\theta)$	Log-partition function and its convex conjugate.
	$D_A(\phi, \theta)$	Bregman divergence induced by the function A.
Fisher information	$I_{x,z}(\theta)$	Fisher information matrix of the distribution $p(x, z \mid \theta)$.
	$I_{z \mid x}(\theta)$	Fisher information matrix of the distribution $p(z \mid x, \theta)$.
Optimization	$t = 1, \ldots, T$	Iteration counter and total iterations.

Acronyms:

MLE	maximum likelihood estimate	GD	gradient descent
MAP	maximum a posteriori estimate	FIM	Fisher information matrix
NLL	negative log-likelihood	KL	Kullback-Leibler
EM	expectation-maximization		

A Supplementary material for Section 2: Expectation-Maximization and Exponential Families

This section extends on the background given in Section 2 and give additional details and properties on

A.1 Expectation-Maximization

This section gives additional details on the derivation of the EM surrogate and some of the perspective taken on the algorithm in the literature. Lange et al. (2000) and Mairal (2013) view EM as a majorization-minimization algorithm to develop a general analysis and extend it to other problems. Chrétien and Hero (2000) and Tseng (2004) view it instead as a proximal point method in Kullback-Leibler divergence to study its asymptotic convergence properties. Finally, Csiszár and Tusnády (1984) and Neal and Hinton (1998) take an alternating minimization procedure view of the algorithm. Csiszár and Tusnády use it to analyze its convergence properties while Neal and Hinton develop an incremental variant. This last perspective is the one presented by Wainwright and Jordan (2008), viewed as a variational method.

The form of the algorithm presented in the main text is the one used by Dempster et al. (1977). The negative log-likelihood (NLL) $\mathcal{L}(\phi)$, surrogate $Q_\theta(\phi)$ and entropy term $H_\theta(\phi)$ are defined as

$$\mathcal{L}(\theta) = -\log p(x \,|\, \theta), \qquad Q_\theta(\phi) = -\int \log p(x, z \,|\, \phi)\, p(z \,|\, x, \theta)\, \mathrm{d}z, \qquad H_\theta(\phi) = -\int \log p(z \,|\, x, \phi)\, p(z \,|\, x, \theta)\, \mathrm{d}z.$$

They obey the decomposition $Q_\theta(\phi) = \mathcal{L}(\phi) + H_\theta(\phi)$. To show this, we use the fact that $\int p(z \,|\, x, \theta)\, \mathrm{d}z = 1$, and

$$\mathcal{L}(\phi) = -\log p(x \,|\, \phi) = -\log p(x \,|\, \phi) \cdot \int p(z \,|\, x, \theta)\, \mathrm{d}z = -\int \log p(x \,|\, \phi)\, p(z \,|\, x, \theta)\, \mathrm{d}z.$$

Along with the chain rule, $p(x, z \,|\, \phi) = p(z \,|\, x, \phi)\, p(x \,|\, \phi)$, we get

$$\begin{aligned}\mathcal{L}(\phi) &= -\int \log p(x \,|\, \phi)\, p(z \,|\, x, \theta)\, \mathrm{d}z \\ &= -\int \log\left(\frac{p(x, z \,|\, \phi)}{p(z \,|\, x, \phi)}\right) p(z \,|\, x, \theta)\, \mathrm{d}z = \overbrace{-\int \log p(x, z \,|\, \phi)\, p(z \,|\, x, \theta)\, \mathrm{d}z}^{Q_\theta(\phi)} + \overbrace{\int \log p(z \,|\, x, \phi)\, p(z \,|\, x, \theta)\, \mathrm{d}z}^{-H_\theta(\phi)}\end{aligned}$$

From a Majorization-Minimization perspective

A majorization-minimization procedure in the sense of Lange et al. (2000) is an iterative procedure to optimize the objective $\mathcal{L}$. Given the current estimate of the parameters θ_t, we first find a majorant, an upper bound f_t that it is tight at θ_t, $\mathcal{L}(\phi) \leq f_t(\phi)$ and $\mathcal{L}(\theta_t) = f_t(\theta_t)$. We then minimize f_t to obtain the new estimate θ_{t+1}. As f_t is an upper bound on the objective, θ_{t+1} is guaranteed to be an improvement if it is an improvement on f_t.

The typical derivation of EM in this setting involves expressing the NLL as the marginal of the complete-data likelihood, multipliying the integrand by $\frac{p(z \,|\, x, \theta)}{p(z \,|\, x, \theta)}$ and using Jensen's inequality, $-\log(\mathbb{E}[x]) \leq -\mathbb{E}[\log(x)]$,

$$\begin{aligned}\mathcal{L}(\phi) &= -\log \int p(x, z \,|\, \phi)\, \mathrm{d}z \\ &= -\log \int p(z \,|\, x, \theta) \frac{p(x, z \,|\, \phi)}{p(z \,|\, x, \theta)}\, \mathrm{d}z \\ &\leq -\int \log\left(\frac{p(x, z \,|\, \phi)}{p(z \,|\, x, \theta)}\right) p(z \,|\, x, \theta)\, \mathrm{d}z = \overbrace{-\int \log p(x, z \,|\, \phi)\, p(z \,|\, x, \theta)\, \mathrm{d}z}^{Q_\theta(\phi)} + \overbrace{\int \log p(z \,|\, x, \theta)\, p(z \,|\, x, \theta)\, \mathrm{d}z}^{-H_\theta(\theta)}.\end{aligned}$$

It gives that the surrogate $Q_\theta(\cdot)$ is an upper bound on the objective, up to a constant, $\mathcal{L}(\phi) \leq Q_\theta(\phi) + \text{const}$. The surrogate $Q_\theta(\cdot)$ itself is not a majorant, as $Q_\theta(\theta) = \mathcal{L}(\theta) + H_\theta(\theta)$. The difference, however, is not relevant for optimization as it does not depend on ϕ. If we define instead the surrogate as $Q'_\theta(\phi) = Q_\theta(\phi) - H_\theta(\theta)$, we get

$$Q'_\theta(\phi) = \mathcal{L}(\phi) + H_\theta(\phi) - H_\theta(\theta). \qquad \text{and} \qquad \mathcal{L}(\theta) = Q'_\theta(\theta)$$

The two formulations of the surrogate share the same minimizers as they differ by an additive constant.

From a proximal point perspective

The definition of $Q'_\theta(\cdot)$ also gives the proximal point perspective used by Chrétien and Hero (2000) and Tseng (2004) to discuss the asymptotic convergence properties of EM. The differences of entropy terms is a KL divergence;

$$\begin{aligned} H_\theta(\phi) - H_\theta(\theta) &= -\int \log p(z \,|\, x, \phi)\, p(z \,|\, x, \theta)\, \mathrm{d}z + \int \log p(z \,|\, x, \theta)\, p(z \,|\, x, \theta)\, \mathrm{d}z, \\ &= -\int \log\left(\frac{p(z \,|\, x, \phi)}{p(z \,|\, x, \theta)}\right) p(z \,|\, x, \theta)\, \mathrm{d}z \quad = \mathrm{KL}[p(z \,|\, x, \theta) \| p(z \,|\, x, \phi)]. \end{aligned}$$

The EM iterations can then be expressed as minimizing $\mathcal{L}$ and a KL proximity term,

$$\theta_{t+1} = \arg\min_\theta \{\mathcal{L}(\theta) + \mathrm{KL}[p(z \,|\, x, \theta) \| p(z \,|\, x, \phi)]\},$$

From an alternating minimization perspective

The expression in terms of a KL divergence also gives the alternating minimization approach used by Csiszár and Tusnády (1984) to show asymptotic convergence, and by Neal and Hinton (1998) to justify partial updates. This is the variational approach presented by Wainwright and Jordan (2008). For a distribution q on the latent variables, parametrized by ϕ, the objective function is equivalent to

$$\mathcal{L}(\theta) = -\log p(x \,|\, \theta) = -\log p(x \,|\, \theta) + \min_\phi \mathrm{KL}[p(z \,|\, x, \theta) \| q(z \,|\, \phi)]$$

if q is sufficiently expressive and we can minimize the KL divergence exactly. The parameters ϕ and θ need not be defined on the same space, as ϕ only controls the conditional distribution over the latent variables and θ controls the complete-data distribution. We can write the EM algorithm as alternating optimization on the augmented objective function

$$\mathcal{L}^+(\theta, \phi) = -\log p(x \,|\, \theta) + \mathrm{KL}[p(z \,|\, x, \theta) \| p(z \,|\, x, \phi)] \qquad \text{such that} \qquad \mathcal{L}(\theta) = \min_\phi \mathcal{L}^+(\theta, \phi).$$

The E and M steps then correspond to

$$\text{E-step:} \qquad \phi_{t+1} = \arg\min_\phi \mathcal{L}^+(\theta_t, \phi), \qquad\qquad \text{M-step:} \qquad \theta_{t+1} = \arg\min_\theta \mathcal{L}^+(\theta, \phi_{t+1}).$$

We will return to this perspective in Appendix E to analyse the progress of the E-step.

Gradients and Hessians

From the equivalence between $Q_\theta(\phi)$ and $Q'_\theta(\phi)$ up to constants, they share the same gradient as the NLL at θ, as

$$\nabla Q_\theta(\theta) = \nabla Q'_\theta(\theta) = \nabla \mathcal{L}(\theta) + \underbrace{\nabla_\phi \mathrm{KL}[p(z \,|\, x, \theta) \| p(z \,|\, x, \phi)] \,|_{\,\phi=\theta}}_{=0},$$

if they are differentiable. Similarly, their Hessian is

$$\nabla^2 Q_\theta(\theta) = \nabla^2 Q'_\theta(\theta) = \nabla^2 \mathcal{L}(\theta) + \nabla^2_\phi \mathrm{KL}[p(z \,|\, x, \theta) \| p(z \,|\, x, \phi)] \,|_{\,\phi=\theta}.$$

Invariance to homeomorphisms

The invariance of the EM update to homeomorphisms is a direct result of the exactness of the M-step. A homeomorphism between two parametrizations (θ, μ) is a continuous bijection f with continous inverse f^{-1}, such that $\theta = f(\mu)$ and $\mu = f^{-1}(\theta)$. Although we use the same notation as the mean and natural parameters, θ and μ can be any parametrization. Letting (θ_t, μ_t) be the current iterates, the EM update in parameters θ or μ yields

$$\theta_{t+1} \in \arg\min_\theta Q_{\theta_t}(\theta) \qquad\qquad \mu_{t+1} \in \arg\min_\mu Q_{f(\mu_t)}(f(\mu)).$$

If $Q_\theta(\cdot)$ is strictly convex, it has a unique minimum and $\theta_{t+1} = f(\mu_{t+1})$, $\mu_{t+1} = f^{-1}(\theta_{t+1})$. Otherwise, (f, f^{-1}) defines a bijection between the possible updates. While the update in some parametrizations might be easier to implement, the update to the probabilistic model is the same regardless of the parametrization.

A.2 Exponential families

For a detailed introduction on exponential families, we recommend the work of Wainwright and Jordan (2008).

An distribution $p(x \,|\, \theta)$ is in the exponential family with natural parameters θ if it has the form

$$p(x \,|\, \theta) = h(x) \exp(\langle S(x), \theta \rangle - A(\theta)) \qquad\qquad -\log p(x \,|\, \theta) = A(\theta) - \langle S(x), \theta \rangle - \log h(x),$$

where h is the base measure, S are the sufficient statistics, and A is the log-parition function. We did not discuss the base measure h in the main text; it is necessary to define the distribution but does not influence the optimization as it does not depend on θ. This can be seen from the gradient and Hessian of the NLL;

$$\nabla -\log p(x \,|\, \theta) = \nabla A(\theta) - S(x) \qquad \text{and} \qquad \nabla^2 -\log p(x \,|\, \theta) = \nabla^2 A(\theta).$$

Examples: Bernoulli and univariate Gaussian

For a binary $x \in \{0, 1\}$, the Bernoulli distribution $p(x \,|\, \pi) = \pi^x (1-\pi)^x$ is an exponential family distribution with

$$h(x) = 1 \qquad S(x) = x \qquad \theta = \log\left(\frac{\pi}{1-\pi}\right) \qquad A(\theta) = \log(1 + e^\theta) = -\log(1-\pi).$$

For $x \in \mathbb{R}$, the Gaussian $p(x \,|\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(x-\mu)^2/2\sigma^2)$ is an exponential family distribution with

$$h(x) = \frac{1}{\sqrt{2\pi}} \qquad S(x) = \left[x, x^2\right] \qquad \theta = \left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right] \qquad A(\theta) = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log\left|-\frac{1}{2\theta_2}\right| = \frac{\mu^2}{2\sigma^2} + \log \sigma.$$

The log-partition function and mean parameters

Given the base measure h and sufficient statistics function S, the log-partition function A is defined such that the probability distribution is valid and integrates to 1,

$$\begin{aligned} 1 = \int p(x \,|\, \theta) \,\mathrm{d}x &= \int h(x) \exp(\langle S(x), \theta \rangle - A(\theta)) \,\mathrm{d}x, \\ &= \exp(-A(\theta)) \int h(x) \exp(\langle S(x), \theta \rangle) \,\mathrm{d}x \end{aligned} \qquad \Longrightarrow \qquad A(\theta) = \log \int h(x) \exp(\langle S(x), \theta \rangle) \,\mathrm{d}x$$

This formulation gives that the log-partition function is convex and its gradient yields the expected sufficient statistics produced by the model, $\nabla A(\theta) = \mathbb{E}_{p(x \,|\, \theta)}[S(x)]$

$$\begin{aligned} \nabla A(\theta) = \nabla \log \int h(x) \exp(\langle S(x), \theta \rangle) \,\mathrm{d}x &= \frac{1}{\int h(x) \exp(\langle S(x), \theta \rangle) \,\mathrm{d}x} \nabla \int h(x) \exp(\langle S(x), \theta \rangle) \,\mathrm{d}x, \\ &= \exp(-A(\theta)) \nabla \int h(x) \exp(\langle S(x), \theta \rangle) \,\mathrm{d}x = \exp(-A(\theta)) \int h(x) \exp S(x) (\langle S(x), \theta \rangle) \,\mathrm{d}x = \int S(x)\, p(x \,|\, \theta) \,\mathrm{d}x. \end{aligned}$$

If the log-partition function A is strictly convex, the exponential family is said to be minimal and there is a bijection between θ and the expected sufficient statistics. The expected sufficient statistics give an equivalent way to parametrize the model, called the mean parameters, which are denoted μ. The gradient ∇A maps the natural to the mean parameters, $\mu = \nabla A(\theta)$. The inverse mapping is the gradient of the convex conjugate of A,

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

We then get the bijection $\mu = \nabla A(\theta)$ and $\theta = \nabla A^*(\mu)$. The Hessians of A and A^* are also inverses of each other. This can be seen from the fact that the composition $\theta = \nabla A^*(\nabla A(\theta))$ is the identity, and

$$\nabla[\nabla A^*(\nabla A(\theta))] = \nabla^2 A^*(\mu)\, \nabla^2 A(\theta) = I.$$

The minimality of the exponential family, or the strict convexity of A, ensures both $\nabla^2 A$ and $\nabla^2 A^*$ are invertible.

For Expectation-Maximization

When the complete-data distribution $p(x, z \,|\, \theta)$ is in the exponential family, the M-step has a simple expression as the surrogate $Q_\theta(\cdot)$ depends on the data only through the expected sufficient statistics at θ,

$$Q_\theta(\phi) = -\int \log p(x, z \,|\, \phi)\, p(z \,|\, x, \theta) \,\mathrm{d}z = -\left\langle \mathbb{E}_{p(z \,|\, x, \theta)}[S(x, z)], \phi \right\rangle + A(\phi).$$

Writing the expected sufficient statistics as $s(\theta) = \mathbb{E}_{p(z \,|\, x, \theta)}[S(x, z)]$, the gradients of the surrogate and NLL are

$$\nabla Q_\theta(\phi) = -\langle s(\theta), \phi \rangle + A(\phi) \qquad \text{and} \qquad \nabla \mathcal{L}(\theta) = \nabla Q_\theta(\theta) = -\langle s(\theta), \phi \rangle + A(\theta).$$

A.3 Bregman divergences

For an overview of Bregman divergences in clustering algorithms and their relation with exponential families, we recommend the work of Banerjee et al. (2005).

Bregman divergence are a generalization of squared Euclidean distance based on convex functions. For a function h, $D_h(\theta, \phi)$ is the difference between the function at θ and its linearization constructed at ϕ,

$$D_h(\theta, \phi) = h(\theta) - h(\phi) - \langle \nabla h(\phi), \theta - \phi \rangle.$$

This is illustrated in Figure 7. The simplest example of a Bregman divergence is the Euclidean distance, which is generated by setting $h(\theta) = \frac{1}{2}\|\theta\|^2$, such that

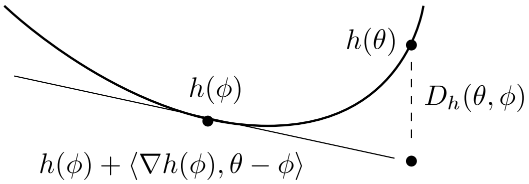

Figure 7: Illustration of the Bregman divergence of a convex function h as the difference between the linearization of the function and its value.

$$\begin{aligned} D_h(\theta, \phi) &= h(\theta) \quad - h(\phi) \quad - \langle \nabla h(\phi), \theta - \phi \rangle, \\ &= \frac{1}{2}\|\theta\|^2 - \frac{1}{2}\|\phi\|^2 - \langle \phi, \theta - \phi \rangle \qquad = \frac{1}{2}\|\theta\|^2 - \langle \phi, \theta \rangle + \frac{1}{2}\|\phi\|^2 \quad = \frac{1}{2}\|\theta - \phi\|^2. \end{aligned}$$

Other examples of Bregman divergences include

Weighted Euclidean/Mahalanobis distance:	$x \in \mathbb{R}^d$	$h(x) = \frac{1}{2}\langle x, Ax \rangle$	$D_h(x, y) = \frac{1}{2}\|x - y\|_A^2$
Kullback-Leibler divergence on the simplex:	$\pi \in \Delta^d$	$h(\pi) = \sum_{i=1}^d \pi_i \log \pi_i$	$D_h(\tau, \pi) = \sum_{i=1}^d \pi_i \log\left(\frac{\pi_i}{\tau_i}\right)$

General properties

The Euclidean example is not representative of general Bregman divergences, as they lack some properties of metrics. They are not necessarily symmetric (in general, $D_h(\theta, \phi) \neq D_h(\phi, \theta)$) and do not satisfy the triangle inequality. The Bregman divergence is convex in its first argument, as it reduces to $h(\theta)$ and a linear term, but needs not be convex in its second argument. The gradients and Hessian with respect to the first argument are

$$\nabla_\theta D_h(\theta, \phi) = \nabla_\theta [h(\theta) - h(\phi) - \langle \nabla h(\phi), \theta - \phi \rangle] = \nabla h(\theta) - \nabla h(\phi) \qquad \text{and} \qquad \nabla_\theta^2 D_h(\theta, \phi) = \nabla^2 h(\theta).$$

Bregman divergences statisfy a generalization of the Euclidean decomposition, called the three point-property;

$$\begin{aligned} \text{Euclidean:} &\qquad \|a - c\|^2 = \|a - b + b - c\|^2 = \|a - b\|^2 + 2\langle a - b, b - c \rangle + \|b - c\|^2, \\ \text{Bregman divergence:} &\qquad D_h(a, c) = D_h(a, b) + \langle a - b, \nabla h(b) - \nabla h(c) \rangle + D_h(b, c). \end{aligned}$$

This property can be directly verified by expanding $D_h(a, b) = h(a) - h(b) - \langle \nabla h(b), a - b \rangle$,

$$\begin{array}{lllllll} & D_h(a, b) & + & \langle a - b, \nabla h(b) - \nabla h(c) \rangle & + & D_h(b, c) \\ = & h(a) - h(b) - \langle \nabla h(b), a - b \rangle & + & \langle \nabla h(b), a - b \rangle - \langle \nabla h(c), a - b \rangle & + & h(b) - h(c) - \langle \nabla h(c), b - c \rangle \\ = & h(a) & & & & - h(c) - \langle \nabla h(c), a - c \rangle = D_h(a, c) \end{array}$$

The Bregman divergence induced by h and its convex conjugate h^* satisfy the following relation,

$$D_h(x, y) = D_{h^*}(\nabla h(y), \nabla h(x)).$$

The convex conjugate of a function h is $h^*(\mu) = \sup_\theta \{\langle \theta, \mu \rangle - h(\theta)\}$, and if h is strictly convex and differentiable, the supremum is attained at $\mu = \nabla h(\theta)$, creating a mapping from the domain of h to the range of its gradient. The inverse mapping can be found by taking the bi-conjugate (the conjugate of the conjugate), which recovers $h = (h^*)^*$; $h(\theta) = \sup_\mu \{\langle \mu, \theta \rangle - h^*(\mu)\}$, and the supremum is attained at $\theta = \nabla h^*(\mu)$.

For exponential families

For an exponential family $p(x \,|\, \theta)$, the Bregman divergence induced by the log-partition function A is the Kullback-Leibler divergence between the distributions given by the parameters

$$\begin{aligned} \mathrm{KL}[p(x \,|\, \phi) \,\|\, p(x \,|\, \theta)] = \int \log\left(\frac{p(x \,|\, \phi)}{p(x \,|\, \theta)}\right) p(x \,|\, \phi) \, \mathrm{d}z &= \int (\langle S(x), \phi \rangle - A(\phi) - \langle S(x), \theta \rangle + A(\theta)) p(x \,|\, \phi) \, \mathrm{d}z, \\ &= A(\theta) - A(\phi) + \left\langle \mathbb{E}_{p(x \,|\, \phi)}[S(x)], \phi - \theta \right\rangle, \\ &= A(\theta) - A(\phi) + \langle \nabla A(\phi), \phi - \theta \rangle \quad = D_A(\theta, \phi). \end{aligned}$$

A.4 Fisher information matrices

For an introduction to Fisher information in the context of the EM algorithm and its connection to the ratio of missing information, we recommend the work of McLachlan and Krishnan (2007, §3.8–3.9).

For a probability distribution parametrized by θ, $p(x \,|\, \theta)$, the Fisher information is a measure of the information that observing some data x would provide about the parameter θ. The Fisher information matrix (FIM) is

$$I(\theta) \;=\; \nabla^2_\phi \mathrm{KL}[p(x \,|\, \theta) \| p(x \,|\, \phi)] \,|_{\,\phi=\theta} \;=\; \mathbb{E}_{p(x \,|\, \theta)}\big[\nabla^2 - \log p(x \,|\, \theta)\big] \;=\; \mathbb{E}_{p(x \,|\, \theta)}\big[\nabla \log p(x \,|\, \theta)\, \nabla \log p(x \,|\, \theta)^\top\big],$$

where all expressions are equivalent. As we will have to distinguish between the information of different distributions, we define the following notation for the distributions $p(x \,|\, \theta)$, $p(x, z \,|\, \theta)$, and $p(z \,|\, x, \theta)$;

$$I_x(\theta) = \mathop{\mathbb{E}}_{p(x \,|\, \theta)}\big[\nabla^2 - \log p(x \,|\, \theta)\big], \quad I_{x,z}(\theta) = \mathop{\mathbb{E}}_{p(x,z \,|\, \theta)}\big[\nabla^2 - \log p(x, z \,|\, \theta)\big], \quad I_{z \,|\, x}(\theta) = \mathop{\mathbb{E}}_{p(z \,|\, x,\theta)}\big[\nabla^2 - \log p(z \,|\, x, \theta)\big].$$

The first two do not depend on data as x and z are sampled from the probabilistic model. The conditional FIM $I_{z \,|\, x}(\theta)$ depends on the observed data x as the expectation is with respect to $p(z \,|\, x, \theta)$.

The Fisher information depends on the parametrization of the distribution. Let us write $I_{x \,|\, \theta}$ and $I_{x \,|\, \mu}$ for the Fisher information of two equivalent parametrizations, (θ, μ), and (f, f^{-1}) be the homeomorphism such that $\theta = f(\mu)$ and $\mu = f^{-1}(\theta)$. The information matrices obey

$$I_{x \,|\, \mu}(\mu) = \mathrm{J}f(\mu)^\top \, I_{x \,|\, \theta}(\theta) \; \mathrm{J}f(\mu),$$

where $\mathrm{J}f$ is the Jacobian of f. Although we use θ and μ, those parametrizations need not be the natural and mean parametrization for this property to hold. This is shown most easily by using the outer-product form;

$$\begin{aligned} I_{x \,|\, \mu}(\mu) &= \mathbb{E}_{p(x \,|\, f(\mu))}\big[\nabla_\mu \log p(x \,|\, f(\mu))\, \nabla_\mu \log p(x \,|\, f(\mu))^\top\big], \\ &= \mathbb{E}_{p(x \,|\, f(\mu))}\big[\mathrm{J}f(\mu)^\top \, \nabla_\theta \log p(x \,|\, \theta)\, \nabla_\theta \log p(x \,|\, \theta)^\top \, \mathrm{J}f(\mu)\big], \\ &= \mathrm{J}f(\mu)^\top \, \mathbb{E}_{p(x \,|\, \theta)}\big[\nabla \log p(x \,|\, \theta)\, \nabla \log p(x \,|\, \theta)^\top\big] \, \mathrm{J}f(\mu) = \mathrm{J}f(\mu)^\top \, I_{x \,|\, \theta}(\theta) \; \mathrm{J}f(\mu). \end{aligned}$$

For an exponential family distribution $p(x \,|\, \theta)$, the FIM is also equal to the Hessian of the NLL, as

$$I(\theta) \;=\; \mathbb{E}_{p(x \,|\, \theta)}\big[\nabla^2 - \log p(x \,|\, \theta)\big] \;=\; \mathbb{E}_{p(x \,|\, \theta)}\big[\nabla^2 A(\theta)\big] \;=\; \nabla^2 A(\theta).$$

For the natural and mean parameters (θ, μ), applying the reparametrization property to $(\nabla A, \nabla A^*)$ along with the fact that $I(\theta) = \nabla^2 A(\theta)$ and $\nabla A(\theta) = [\nabla A^*(\mu)]^{-1}$ gives that $I_{x \,|\, \mu}(\mu) = I_{x \,|\, \theta}(\theta)^{-1}$, as

$$I_{x \,|\, \mu}(\mu) = \nabla^2 A^*(\mu) I_{x \,|\, \theta}(\theta)\, \nabla^2 A^*(\mu) = \nabla^2 A^*(\mu)\, \nabla^2 A(\theta)\, \nabla^2 A^*(\mu) = \nabla^2 A^*(\mu).$$

For Expectation-Maximization, if the complete-data distribution $p(x, z \,|\, \theta)$ is in the exponential family, the Hessian of the surrogate and objective are

$$\nabla^2 Q_\theta(\theta) = \nabla^2 A(\theta) = I_{x,z}(\theta), \qquad \begin{aligned} \nabla^2 \mathcal{L}(\theta) &= \nabla^2 Q_\theta(\theta) - \nabla^2_\phi \mathrm{KL}[p(z \,|\, x, \theta) \| p(z \,|\, x, \phi)] \,|_{\,\phi=\theta} \\ &= I_{x,z}(\theta) - I_{z \,|\, x}(\theta). \end{aligned}$$

This follows from the definition of $Q_\theta(\cdot)$ (Appendix A.1) and the properties of exponential families (Appendix A.2).

Natural gradients

The gradient is a measure of the direction of steepest increase, where steepest is defined with respect to the Euclidean distance between the parameters. When the parameters of a function also define a probability distribution, the natural gradient (Amari and Nagaoka, 2000) is the direction of steepest increase, where steepest is instead measured by the KL divergence between the induced distributions. The natural gradient is obtained by preconditioning the gradient with the inverse of the FIM of the relevant distribution, $I(\theta)^{-1} \, \nabla\mathcal{L}(\theta)$.

For exponential families, the gradient with respect to the natural parameters θ is the natural gradient with respect to the mean parameters μ. Letting $\mathcal{L}_d(\mu) = \mathcal{L}(\nabla A^*(\mu))$ be the objective express in mean parameters, we have

$$\nabla\mathcal{L}_d(\mu) = \nabla^2 A^*(\mu)\, \nabla\mathcal{L}(\theta) = [\nabla^2 A(\theta)]^{-1}\, \nabla\mathcal{L}(\theta) \qquad \nabla\mathcal{L}(\theta) = \nabla^2 A(\theta)\, \nabla\mathcal{L}_d(\mu) = [\nabla^2 A^*(\mu)]^{-1}\, \nabla\mathcal{L}_d(\mu)$$

This implies the mirror descent update $\mu_{t+1} = \mu_t - \nabla\mathcal{L}(\theta_t)$ is a natural gradient descent step in mean parameters when the mirror map A is the log-partition function of an exponential family (Raskutti and Mukherjee, 2015). The view of EM as a natural gradient update was already used by Sato (1999) to justify a stochastic variant.

A.5 Mirror descent, convexity, smoothness and their relative equivalent

For a more thorough coverage of mirror descent, we recommend the works of Nemirovski and Yudin (1983) and Beck and Teboulle (2003). For an introduction on convexity, smoothness and strong convexity, we recommend the work of Nesterov (2013). For their relative equivalent, see Bauschke et al. (2017) and Lu et al. (2018).

The traditional gradient descent algorithm to optimize a function f can be expressed as the minimization of the linearization of f at the current iterates θ_t and a Euclidean distance proximity term depending on the step-size γ,

$$\theta_{t+1} = \arg\min_\theta \Big\{ f(\theta_t) + \langle \nabla f(\theta_t), \theta - \theta_t \rangle + \tfrac{1}{2\gamma} \|\theta - \theta_{t+1}\|^2 \Big\}.$$

As the surrogate objective is convex, the update is found by taking the derivative and setting it to zero;

$$\nabla f(\theta_t) + \tfrac{1}{\gamma}(\theta_{t+1} - \theta_t) = 0 \qquad \implies \qquad \theta_{t+1} = \theta_t - \gamma \nabla f(\theta_t).$$

The mirror descent algorithm is an extension where the Euclidean distance is replaced by a Bregman divergence,

$$\theta' = \arg\min_\phi \Big\{ f(\theta) + \langle \nabla f(\theta), \phi - \theta \rangle + \tfrac{1}{\gamma} D_h(\phi, \theta) \Big\}.$$

Setting $h(\theta) = \frac{1}{2}\|\theta\|^2$ recovers the gradient descent surrogate. The stationarity condition gives the update

$$\nabla f(\theta_t) + \tfrac{1}{\gamma}(\nabla h(\theta_{t+1}) - \nabla h(\theta_t)) = 0 \qquad \implies \qquad \nabla h(\theta_{t+1}) = \nabla h(\theta_t) - \gamma \nabla f(\theta_t).$$

Or, equivalently, the update can be written in the dual parametrization $\mu = \nabla h(\theta)$,

$$\mu_{t+1} = \mu_t - \gamma \nabla f(\theta_t).$$

The mirror descent update applies the gradient step to the dual parameters instead of the primal parameters θ. In the mirror descent literature, the reference function h is called the mirror function or mirror map.

Smoothness and strong convexity

The gradient descent update with an arbitrary constant step-size γ is not guaranteed to make progress on the original function f, at least not without additional assumptions. A common assumption is that the function f is smooth, meaning that its gradient is Lipschitz with constant L,

$$\|\nabla f(\theta) - \nabla f(\phi)\| \leq L \|\theta - \phi\|, \qquad \text{for any } \theta, \phi.$$

The L-smoothness of f implies the following upper bound holds,

$$f(\phi) \leq f(\theta) + \langle \nabla f(\theta), \phi - \theta) \rangle + \frac{L}{2} \|\theta - \phi\|^2 \qquad \text{for any } \theta, \phi.$$

Setting $\gamma \leq \frac{1}{L}$ ensures the surrogate optimized by gradient descent is an upper bound on f and leads to progress. If the objective function is also α-strongly convex, meaning the following lower bound holds,

$$f(\theta) + \langle \nabla f(\theta), \phi - \theta \rangle - \frac{\alpha}{2} \|\theta - \phi\|^2 \leq f(\phi) \qquad \text{for } \alpha > 0 \text{ and any } \theta, \phi,$$

gradient descent converges at a faster, linear rate. This definition recovers convexity in the case $\alpha = 0$ and is otherwise stronger. If f is twice differentiable, α-strong convexity and L-smoothness are equivalent to

$$\alpha I \preceq \nabla^2 f(\theta) \preceq LI \qquad \text{for all } \theta.$$

Here, $\preceq$ is the Loewner ordering on matrices, where $A \preceq B$ if $B - A$ is positive semi-definite, meaning the minimum eigenvalue of $B - A$ is larger than or equal to zero.

Relative smoothness and strong convexity

Relative L-smoothness and α-strong convexity provide an analog of smoothness and strong-convexity for mirror descent. They are defined relative to a reference function h, such that the following lower and upper bound hold

$$f(\theta) + \langle \nabla f(\theta), \phi - \theta) \rangle - \alpha D_h(\phi, \theta) \leq f(\phi) \leq f(\theta) + \langle \nabla f(\theta), \phi - \theta) \rangle + L D_h(\phi, \theta) \qquad \text{for } 0 < \alpha \leq L \text{ and any } \theta, \phi.$$

Alternatively, if f and h are twice differentiable, those conditions are equivalent to

$$\alpha \nabla^2 h(\theta) \preceq \nabla^2 f(\theta) \preceq L \nabla^2 h(\theta) \qquad \text{for all } \theta.$$

In the case $h(\theta) = \frac{1}{2}\|\theta\|^2$, we recover the standard definition of Euclidean smoothness and strong-convexity.

B Supplementary material for Section 3: EM and Mirror Descent

This section gives additional details on the relationship between EM and mirror descent and the 1-relative smoothness of EM. We restate in longer form the proof of Proposition 1;

PROPOSITION 1. *For exponential family distributions, the M-step update in Expectation-Maximization is equivalent to the minimization of the following upper-bound;*

$$\mathcal{L}(\phi) \leq \mathcal{L}(\theta) + \langle \nabla \mathcal{L}(\theta), \phi - \theta \rangle + D_A(\phi, \theta), \tag{7}$$

where A is the log-partition of the complete-data distribution, and $D_A(\phi, \theta) = \mathrm{KL}[p(x, z \,|\, \theta) \| p(x, z \,|\, \phi)]$.

Proof of Proposition 1. Recall the decomposition of the surrogate in terms of the objective and entropy term, $Q_\theta(\phi) = \mathcal{L}(\phi) + H_\theta(\phi)$ in Equation (2). It gives

$$\mathcal{L}(\phi) - \mathcal{L}(\theta) = Q_\theta(\phi) - Q_\theta(\theta) + H_\theta(\theta) - H_\theta(\phi),$$

where $H_\theta(\theta) - H_\theta(\phi) \leq 0$ as $H_\theta(\phi)$ is minimized at $\phi = \theta$. We will show that for exponential families,

$$Q_\theta(\phi) - Q_\theta(\theta) = \langle \nabla \mathcal{L}(\theta), \phi - \theta \rangle + D_A(\phi, \theta),$$

which implies the upper-bound in Equation (7) and that its minimum matches that of $Q_\theta(\phi)$.

If the complete-data distribution is in the exponential family, the surrogate in natural parameters is

$$\begin{aligned} Q_\theta(\phi) &= -\int \log p(x, z \,|\, \phi)\, p(z \,|\, x, \theta)\, \mathrm{d}z, \\ &= -\int [\langle S(x, z), \phi \rangle - A(\phi)]\, p(z \,|\, x, \theta)\, \mathrm{d}z = -\left\langle \mathbb{E}_{p(z \,|\, x, \theta)}[S(x, z)], \phi \right\rangle + A(\phi). \end{aligned}$$

For simplicity of notation, we write $s(\theta)$ for the expected sufficient statistics $\mathbb{E}_{p(z \,|\, x, \theta)}[S(x, z)]$ (while the $s(\theta)$ depends on x and we could write $s(\theta, x)$, we ignore it as the same x is always given to s). We will use the definition of the Bregman divergence and the fact that the gradient of the surrogate matches the gradient of the objective,

$$D_A(\phi, \theta) = A(\phi) - A(\theta) - \langle \nabla A(\theta), \phi - \theta \rangle, \qquad \text{and} \qquad \nabla \mathcal{L}(\theta) = \nabla Q_\theta(\theta) = \nabla A(\theta) - s(\theta).$$

Expanding $Q_\theta(\phi) - Q_\theta(\theta)$, we have

$$\begin{aligned} Q_\theta(\phi) - Q_\theta(\theta) &= -\langle s(\theta), \phi - \theta \rangle + A(\phi) - A(\theta), \\ &= -\langle s(\theta), \phi - \theta \rangle + \langle \nabla A(\theta), \phi - \theta \rangle + A(\phi) - A(\theta) - \langle \nabla A(\theta), \phi - \theta \rangle, && (\pm \langle \nabla A(\theta), \phi - \theta \rangle) \\ &= -\langle s(\theta) - \nabla A(\theta), \phi - \theta \rangle + D_A(\phi, \theta), && (D_A(\phi, \theta) = A(\phi) - A(\theta) - \langle \nabla A(\theta), \phi - \theta \rangle) \\ &= \langle \nabla \mathcal{L}(\theta), \phi - \theta \rangle + D_A(\phi, \theta). \end{aligned}$$

□

For completeness, we present an alternative derivation that relies on additional material in Appendices A.1–A.5. That the M-step is a mirror descent step can be seen from the stationary point of $Q_\theta(\phi)$ and Equation (7),

$$A(\phi) = s(\theta) \qquad \text{and} \qquad A(\phi) = A(\theta) - \nabla \mathcal{L}(\theta) = s(\theta).$$

To show the upper bound holds, we can use the expansion of the objective as $\mathcal{L}(\phi) = Q_\theta(\phi) - H_\theta(\phi)$ to get

$$\nabla^2 \mathcal{L}(\theta) = \nabla^2 Q_\theta(\phi)\, \nabla^2 \mathrm{KL}[p(z \,|\, \theta) \| p(z \,|\, \phi)] = \nabla^2 A(\phi) - I_{z \,|\, x}(\phi),$$

where $I_{z \,|\, x}(\phi)$ is the FIM of $p(z \,|\, x, \phi)$. As Fisher information matrices are positive semi-definite, we get that $\nabla^2 \mathcal{L}(\theta) \preceq \nabla^2 A(\theta)$, establishing the 1-smoothness of EM relative to A and the upper bound in Equation (7).

Equivalence between stochastic EM and stochastic mirror descent

We now look at variants of EM based on stochastic approximation and show they can be cast as stochastic mirror descent. We focus on the online EM of Cappé and Moulines (2009), but it also applies to the incremental and stochastic versions of Neal and Hinton (1998), Sato (1999), and Delyon et al. (1999).

The stochastic version of the EM update uses only a subset of samples per iteration to compute the E-step and applies the M-step to the average of the sufficient statistics observed so far. We assume we have n independent samples for the observed variables, $x_1, \dots, x_n$, such that the objective factorizes as

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \mathcal{L}_i(\theta) = \sum_{i=1}^{n} \log p(x_i \,|\, \theta)$$

Defining the individual expected sufficient statistics as $s_i(\theta_t) = \mathbb{E}_{p(z \,|\, x_i, \theta_t)}[S(x_i, z)]$, the online EM algorithm updates a running average of sufficient statistics using a step-size γ_t

$$\mu_{t+1} = (1-\gamma_t)\mu_t + \gamma_t s_{i_t}(\theta_t), \qquad \text{where } i_t \sim U[n]. \tag{14}$$

With step-sizes $\gamma_t = 1/t$, the mean parameters at step t are the average of the observed sufficient statistics,

$$\mu_T = (1-\gamma_T)\mu_{T-1} + \gamma_T s_{i_T}(\theta_T) = \frac{T-1}{T}\mu_{T-1} + \frac{1}{T}s_{i_T}(\theta_T) = \frac{1}{T}\sum_{t=1}^{T} s_{i_t}(\theta_t).$$

The natural parameters are then updated with $\theta_t = \nabla A^*(\mu_t)$.

Proposition 4. *The online EM algorithm (Eq. 14) is equivalent to the stochastic mirror descent update*

$$\theta_{t+1} = \arg\min_{\phi} \left\{ \langle \nabla \mathcal{L}_{i_t}(\theta_t), \phi - \theta_t \rangle + \frac{1}{\gamma_t} D_A(\phi, \theta_t) \right\}, \qquad \textit{with } i_t \sim U[n]. \tag{15}$$

Proof of Proposition 4. We show the equivalence of one step, assuming they select the same index i_t. The online EM update (Eq. 14) guarantees the natural and mean parameters match, $\theta_t = \nabla A^*(\mu_t)$, and the update to θ_{t+1} is

$$\theta_{t+1} = \nabla A^*\big((1-\gamma_t)\mu_t + \gamma_t s_{i_t}(\theta_t)\big).$$

where $s_i(\theta) = \mathbb{E}_{p(z \,|\, x_i, \theta)}[S(x_i, z)]$. The stationary point of Equation (15), on the other hand, ensures

$$0 = \nabla \mathcal{L}_{i_t}(\theta_t) + \frac{1}{\gamma_t}(\nabla A(\theta_{t+1}) - \nabla A(\theta_t)) \quad \implies \quad \theta_{t+1} = \nabla A^*(\nabla A(\theta_t) - \gamma_t \nabla \mathcal{L}_{i_t}(\theta_t)).$$

As in the proof of Proposition 1 (Appendix B), using that the gradient of the loss and the surrogate match,

$$\nabla \mathcal{L}_{i_t}(\theta_t) = -s_{i_t}(\theta_t) + \nabla A(\theta_t),$$

we get that both update match,

$$\nabla A(\theta_t) - \gamma_t \nabla \mathcal{L}_{i_t}(\theta_t)(1-\gamma_t)\nabla A(\theta_t) + \gamma_t s_{i_t}(\theta_t) \qquad \implies \qquad \mu_{t+1} = (1-\gamma_t)\mu_t + \gamma_t s_{i_t}(\theta_t). \qquad \square$$

C Supplementary material for Section 4: Assumptions and Open Constraints

This section gives additional details on the assumptions discussed in Section 4 and shows the derivation for maximum a posteriori (MAP) estimation with EM under a conjugate prior. We first mention how **A3** implies that the EM iterates are well-defined and introduce notation to discuss proper conjugate priors for exponential families. We then show that a proper prior implies that the surrogate optimized by EM leads to well-defined solutions and satisfies **A3**, and end with showing how an equivalent of Proposition 2 holds for MAP.

A3 makes the update are well defined. Consider fitting the variance $\sigma^2 > 0$ of a Gaussian. The possible pitfalls for the update is to be on the boundary ($\sigma^2 = 0$), or diverge ($\sigma^2 \to \infty$). Here is how **A3** avoids those cases.

The exponential family assumptions imply that the surrogate $Q_{\theta_t}(\cdot)$ optimized during the M-step is convex. Its domain, Ω, is an open set. **A3** constrains the sub-level sets $\Omega_{\theta_t} = \{\phi \in \Omega : Q_{\theta_t}(\phi) \leq Q_{\theta_t}(\theta)\}$ to be compact (closed and bounded). As the minimum of the surrogate is contained in any sub-level set, it must be finite (as the sub-level sets are bounded) and contained strictly in Ω (as the sub-level sets are closed).

Proper conjugate priors. We first discuss exponential families, without the added complexity of EM. In the main text, we used x to denote the entire dataset. To discuss priors, it is useful to consider the dataset as n i.i.d. observations $x_1, \dots, x_n$ from a (minimal, regular) exponential family, with negative log-likelihood (NLL)

$$p(x_i \,|\, \theta) \propto \exp(\langle T(x_i), \theta\rangle - A(\theta)), \qquad \text{NLL}(\theta) = -\sum_{i=1}^n \log p(x_i \,|\, \theta) = \sum_{i=1}^n A(\theta) - \langle T(x_i), \theta\rangle.$$

For exponential families, parametrizing the prior by a strength $n_0 > 0$ and the sufficient statistics m_0 we expect to observe a priori, the conjugate prior that leads to the same form for the posterior is

$$p(\theta \,|\, m_0, n_0) \propto \exp(\langle m_0, \theta\rangle - n_0 A(\theta)).$$

The regularized objective of adding the NLL and the prior is then, up to a multiplicative constant of $n + n_0$,

$$\mathcal{L}(\theta) = \frac{1}{n+n_0}\left(-\sum_{i=1}^n \log p(x_i \,|\, \theta) - \log p(\theta \,|\, m_0, n_0)\right) = A(\theta) - \langle \bar{m}, \theta\rangle, \quad \text{with } \bar{m} = \frac{m_0 + \sum_{i=1}^n T(x_i)}{n+n_0}. \tag{16}$$

To discuss proper priors, we need to discuss the constraint set Ω in more details. For a d-dimensional, regular, minimal exponential family, the set of valid natural parameters is defined from the log-partition function as $\Omega = \{\theta \in \mathbb{R}^d \,|\, A(\theta) < \infty\}$. The equivalent set of mean parameters, through the bijection $(\nabla A, \nabla A^*)$, is

$$\mathcal{M} = \{\mu \in \mathbb{R}^d \,|\, \exists \theta \in \Omega : \mathbb{E}_{p(x \,|\, \theta)}[T(x)] = \mu\} \quad (\text{or } \mathcal{M} = \nabla A(\Omega)),$$

the image of Ω through ∇A (Wainwright and Jordan, 2008, Theorem 3.3). For the prior to be proper, the expected sufficient statistics under the prior m_0 need to be in the interior of $\mathcal{M}$ (Diaconis and Ylvisaker, 1979, Theorem 1).

MAP solutions are well defined. The sufficient statistics $T(x_i)$ could lie on the boundary of $\mathcal{M}$, which is why the MLE is sometimes ill-defined. For example, estimating the covariance of a Gaussian from one sample leads to $\sigma^2 = 0$. However, if the prior is proper, $m_0 \in \mathcal{M}$ then the average $\bar{m} = \frac{1}{n+n_0}(\sum_{i=1}^n T(x_i) + m_0)$ will also be in $\mathcal{M}$. By convexity, the MAP is at the stationary point of Equation (16), $\nabla A(\bar{\theta}) = \bar{m}$ and $\bar{\theta} = \nabla A^*(\bar{m})$ will be in Ω.

For completeness, let us show that this also implies **A3**. As Ω is convex, it is sufficient to show that $\mathcal{L}(\theta) \to \infty$ from any direction $v \in \mathbb{R}^d$ starting from $\bar{\theta}$, leading to the sequence $\theta(t) = \bar{\theta} + tv$ for $t > 0$.

- If $\theta(t)$ crosses the boundary of Ω, $\mathcal{L}(\theta(t)) \to \infty$ due to the log-partition function.

  Let t_b be the finite crossing point. The parameters $\theta(t_b)$ and the inner product $\langle \bar{m}, \theta(t_b)\rangle$ are also finite. But since the boundary of Ω is defined by $A(\theta) < \infty$, by (lower-semi-)continuity of A, $\lim_{t \to t_b^-} A(\theta(t)) = \infty$.

- If $\theta(t)$ does not cross a boundary, $\mathcal{L}(\theta(t)) \to \infty$ by strict convexity.

  Consider the restriction of $\mathcal{L}$ to the line spanned by v, $f(t) = \mathcal{L}(\theta(t))$ for $t > 0$. By the properties of $\mathcal{L}$, $f(t)$ is strictly convex and minimized at $t = 0$. Let $t_0 > 0$ be an arbitrary point. By strict convexity,

  $$f(t) > f(t_0) + f'(t_0)(t - t_0) \qquad \text{and} \qquad f'(t_0) > 0$$

  for some finite $f(t_0)$ and $f'(t_0)$. Taking the limit of the lower bound as $t \to \infty$ gives that $\lim_{t\to\infty} \mathcal{L}(\theta(t)) = \infty$.

For background on the constraint sets of parameters exponential families, we recommend Wainwright and Jordan (2008, §3.4). For a geometric view on priors in Bregman divergences, see Agarwal and Daumé III (2010).

EM with a prior. We now consider the analysis of EM with a proper conjugate prior if the full-data distribution $p(x, z \,|\, \theta)$ is in the exponential family. Assuming that the observed and latent variables can be partitioned into i.i.d. pairs (x_i, z_i), as is the case for example with Gaussian mixture models, the likelihood for a full observation is

$$p(x_i, z_i \,|\, \theta) \propto \exp(\langle T(x_i, z_i), \theta\rangle - A(\theta)).$$

A conjugate prior on θ will have the same form as above,

$$p(\theta \,|\, m_0, n_0) \propto \exp(\langle m_0, \theta\rangle - n_0 A(\theta)),$$

and the MAP–EM objective will have the form (up to the normalization constant $n + n_0$)

$$\mathcal{L}_{\text{MAP}}(\theta) = -\frac{1}{n+n_0}\left(\sum_{i=1}^{n} \log p(x \,|\, \theta) - \log p(\theta \,|\, m_0, n_0)\right).$$

Applying the same upper bounds as in the MLE case, we can define the MAP–EM surrogate as

$$\begin{aligned}\tilde{Q}_\theta(\phi) &= \frac{1}{n+n_0}\left(\sum_{i=1}^{n}\int \log p(x_i \,|\, z_i, \phi) p(z_i \,|\, x_i, \theta)\,\mathrm{d}z + n_0 A(\theta) - \langle m_0, \theta\rangle\right),\\ &= \frac{1}{n+n_0}\left(\textstyle\sum_{i=1}^{n} A(\phi) - \left\langle \mathbb{E}_{p(z_i \,|\, x_i, \theta)}[T(x_i, z_i)], \theta\right\rangle + n_0 A(\phi) - \langle m_0, \phi\rangle\right).\end{aligned}$$

Writing $\bar{s}(\theta) = \sum_{i=1}^{n} \mathbb{E}_{p(z_i \,|\, x_i, \theta)}[T(x_i, z_i)]$ for the sum of sufficient statistics, the surrogate is

$$= A(\phi) - \langle \bar{m}(\theta), \phi\rangle, \qquad \text{where } \bar{m}(\theta) = \frac{\bar{s}(\theta) + m_0}{n + n_0}.$$

Ignoring the rescaling by $n + n_0$, this only changes the original surrogate by adding a linear term. The rescaled objective is still 1-smooth[4] relative to A, and the results derived for MLE still hold for MAP, up to minor variations. Writing $\mathcal{L}$ and $\mathcal{L}_{\text{MAP}}$ for the non-regularized MLE and regularized MAP objectives, the equivalent of Proposition 2 includes the prior in the optimality gap;

Proposition 5. *Under assumptions **A1**–**A3**, EM for exponential family distributions with a proper conjugate prior $p(\theta \,|\, m_0, n_0) \propto \exp(\langle m_0, \theta\rangle - n_0 A(\theta))$ converges at the rate*

$$\begin{aligned}\min_{t \le T} \mathrm{KL}[p(x, z \,|\, \theta_{t+1}) \| p(x, z \,|\, \theta_t)] &\le \frac{\mathcal{L}_{\text{MAP}}(\theta_1) - \mathcal{L}_{\text{MAP}}(\theta^*)}{T} \qquad \text{(where } \theta^* \text{ is a minimum of } \mathcal{L}_{\text{MAP}})\\ &= \frac{\mathcal{L}(\theta_1) - \mathcal{L}(\theta^*)}{T} + \frac{\log p(\theta^* \,|\, m_0, n_0) - \log p(\theta_1 \,|\, m_0, n_0)}{T}.\end{aligned}$$

The proof follows the same steps as Proposition 2, and similar variants hold for the locally convex (Corollary 1) and strongly-convex (Corollary 3) cases. To relate the convergence of the successive iterates of Proposition 5 to stationarity, a similar development as for Corollary 2 with the notation introduced above gives

Corollary 4. *Under assumptions **A1**–**A3**, with a proper conjugate prior $p(\theta \,|\, m_0, n_0) \propto \exp(\langle m_0, \theta\rangle - n_0 A(\theta))$,*

$$\min_{t \le T} D_{A^*}\left(\frac{\bar{s}(\theta_t) + m_0}{n + n_0}, \mu_t\right) \le \frac{\mathcal{L}_{\text{MAP}}(\theta_1) - \mathcal{L}_{\text{MAP}}(\theta^*)}{T}.$$

The average of the prior and observed sufficient statistics $\frac{\bar{s}(\theta_t)+m_0}{n+n_0}$ and the mean parameters μ_t are the two parts of the regularized gradient, $\nabla\mathcal{L}_{\text{MAP}}(\theta_t) = \nabla A(\theta_t) - \frac{\bar{s}(\theta_t)+m_0}{n+n_0}$, and $D_{A^*}\left(\frac{\bar{s}(\theta_t)+m_0}{n+n_0}, \mu_t\right) = 0$ implies $\nabla\mathcal{L}_{\text{MAP}}(\theta_t) = 0$.

[4] Without rescaling, the MLE and MAP objectives would be n-smooth and $(n + n_0)$-smooth relative to A. While rescaling changes the constants, the resulting algorithm is the same; running GD with step-size γ on a function f is equivalent to a step-size γ/C on $f' = Cf$.

D Supplementary material for Section 5: Convergence of EM for Exponential Families

This section presents additional details and proofs for the results in Section 5;

D.1 Convergence of EM to stationary points (Proposition 2 and Corollary 2)

Proposition 2. *Under assumptions* ***A1–A3****, EM for exponential family distributions converges at the rate*

$$\min_{t\leq T} \mathrm{KL}[p(x,z\,|\,\theta_{t+1})\|p(x,z\,|\,\theta_t)] \leq \frac{\mathcal{L}(\theta_1)-\mathcal{L}^*}{T}.$$

Proof of Proposition 2. Assumptions **A1–A3** ensure that the updates are well defined. **A1** ensures the mapping $(\nabla A, \nabla A^*)$ is well defined and the update $\theta_t \to \theta_{t+1}$ is unique. **A2** ensures the objective is lower-bounded by some value $\mathcal{L}^*$ and **A3** ensures that, if the parameters are restricted to an open set Ω, the updates remain in Ω as long as $\theta_1 \in \Omega$. Proposition 1 then gives that a step from θ_t to θ_{t+1} satisfies

$$\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) + \langle \nabla\mathcal{L}(\theta_t), \theta_{t+1}-\theta_t\rangle + D_A(\theta_{t+1},\theta_t).$$

As θ_{t+1} is selected to minimize the upper bound, it is at a stationary point. Using that $\nabla D_A(\theta,\theta_t) = \nabla A(\theta) - \nabla A(\theta_t)$,

$$\nabla_{\theta_{t+1}}\{\langle \nabla\mathcal{L}(\theta_t), \theta_{t+1}-\theta_t\rangle + D_A(\theta_{t+1},\theta_t)\} = 0 \quad\implies\quad \nabla\mathcal{L}(\theta_t) + \nabla A(\theta_{t+1}) - \nabla A(\theta_t) = 0.$$

Substituting $\nabla\mathcal{L}(\theta_t)$ for $\nabla A(\theta_t) - \nabla A(\theta_{t+1})$ in the upper bound and using the definition of Bregman divergences

$$D_A(\theta_{t+1},\theta_t) = A(\theta_{t+1}) - A(\theta_t) - \langle \nabla A(\theta_t), \theta_{t+1}-\theta_t\rangle,$$

gives the simplification

$$\begin{aligned}\mathcal{L}(\theta_{t+1}) &\leq \mathcal{L}(\theta_t) + \langle \nabla\mathcal{L}(\theta_t), \theta_{t+1}-\theta_t\rangle + D_A(\theta_{t+1},\theta_t)\\ &= \mathcal{L}(\theta_t) + \langle \nabla A(\theta_t) - \nabla A(\theta_{t+1}), \theta_{t+1}-\theta_t\rangle + D_A(\theta_{t+1},\theta_t),\\ &= \mathcal{L}(\theta_t) - \langle \nabla A(\theta_{t+1}), \theta_{t+1}-\theta_t\rangle + A(\theta_{t+1}) - A(\theta_t) \quad = \mathcal{L}(\theta_t) - D_A(\theta_t,\theta_{t+1}).\end{aligned}$$

Reorganizing the inequality, we have that

$$D_A(\theta_t,\theta_{t+1}) \leq \mathcal{L}(\theta_t) - \mathcal{L}(\theta_{t+1}).$$

Summing over all iterations $t = 1,\dots,T$ and dividing by T gives the result,

$$\min_{t\leq T} D_A(\theta_t,\theta_{t+1}) \leq \frac{1}{T}\sum_{t=1}^{T} D_A(\theta_t,\theta_{t+1}) \leq \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(\theta_t) - \mathcal{L}(\theta_{t+1}) = \frac{\mathcal{L}(\theta_1)-\mathcal{L}(\theta_T)}{T}.$$

Using the lower-bound on the objective function, $\mathcal{L}(\theta_T) \geq \mathcal{L}^*$, finishes the proof. □

Corollary 2. *Under assumptions* ***A1–A3****,*

$$\min_{t\leq T} D_{A^*}(s(\theta_t),\mu_t) \leq \frac{\mathcal{L}(\theta_1)-\mathcal{L}^*}{T}.$$

The observed sufficient statistics $s(\theta_t)$ and mean parameters μ_t are the two parts of the gradient, $\nabla\mathcal{L}(\theta_t) = \mu_t - s(\theta_t)$, and $D_{A^*}(s(\theta_t),\mu_t) = 0$ implies $\nabla\mathcal{L}(\theta_t) = 0$.

Proof of Corollary 2. The proof follows from Proposition 2 and the form of the update. We have that

$$\begin{aligned}
&\text{the update ensures} && \nabla\mathcal{L}(\theta_t) = \nabla A(\theta_t) - \nabla A(\theta_{t+1}),\\
&\text{the gradient is} && \nabla\mathcal{L}(\theta_t) = \nabla A(\theta_t) - s(\theta_t),\\
&\text{the Bregman divergence satisfies} && D_A(\theta_t,\theta_{t+1}) = D_{A^*}(\nabla A(\theta_{t+1}), \nabla A(\theta_t)).
\end{aligned}$$

Using the mapping between natural and mean parameters, we get

$$D_A(\theta_t,\theta_{t+1}) = D_{A^*}(\mu_{t+1},\mu_t) = D_{A^*}(\mu_t - \nabla\mathcal{L}(\theta_t),\mu_t) = D_{A^*}(s(\theta_t),\mu_t). \qquad \square$$

D.2 Natural decrement (Section 5.1)

For a small perturbation δ, the Bregman divergence is well approximated by its second-order Taylor expansion

$$D_{A^*}(\mu+\delta,\mu) = \underbrace{D_{A^*}(\mu,\mu)}_{=0} + \langle \underbrace{\nabla_{\mu'} D_{A^*}(\mu',\mu)\,|_{\,\mu'=\mu}}_{=0}, \delta\rangle + \frac{1}{2}\langle \delta, \underbrace{\nabla^2_{\mu'} D_{A^*}(\mu',\mu)\,|_{\,\mu'=\mu}}_{\nabla^2 A^*(\mu)}\,\delta\rangle + o(\|\delta\|^3) \approx \frac{1}{2}\|\delta\|^2_{\nabla^2 A^*(\mu)}.$$

Using that $\nabla^2 A^*(\mu) = [\nabla^2 A(\theta)]^{-1} = I_{x,z}(\theta)^{-1}$ (see Appendix A.4) and $\delta = \nabla\mathcal{L}(\theta)$, we get an Euclidean approximation of what the divergence measures, which we call the "natural decrement" as a reference to the Newton decrement used in the affine-invariant analysis of Newton's method (Nesterov and Nemirovski, 1994)

$$\text{natural decrement:} \quad \frac{1}{2}\|\nabla\mathcal{L}(\theta)\|^2_{I_{x,z}(\theta)^{-1}} \qquad\qquad \text{Newton decrement:} \quad \frac{1}{2}\|\nabla\mathcal{L}(\theta)\|^2_{\nabla^2\mathcal{L}(\theta)^{-1}}$$

The invariance to homeomorphisms can be shown as follow. Consider an alternative parametrization of the objective, $\mathcal{L}_{\text{alt}}(\psi) = \mathcal{L}(f(\psi))$ where (f, f^{-1}) is the mapping between the parametrizations, $\theta = f(\psi)$ and $\psi = f^{-1}(\theta)$. We use $I_{x,z\,|\,\theta}$ and $I_{x,z\,|\,\psi}$ to differentiate between the FIM of the two parametrizations. We have

$$\nabla\mathcal{L}_{\text{alt}}(\psi) = \nabla\mathcal{L}(f(\psi)) = \mathrm{J}f(\psi)\,\nabla\mathcal{L}(\theta) \qquad \text{and} \qquad I_{x,z\,|\,\psi}(\psi) = \mathrm{J}f(\psi)^\top\, I_{x,z\,|\,\theta}(\theta)\ \mathrm{J}f(\psi),$$

where the second equality is a property of the Fisher information, shown in Appendix A.4. The two parametrizations then give the same natural decrement,

$$\begin{aligned}
\|\nabla\mathcal{L}_{\text{alt}}(\psi)\|^2_{I_{x,z\,|\,\psi}(\psi)^{-1}} &= \left\langle \nabla\mathcal{L}_{\text{alt}}(\psi), I_{x,z\,|\,\psi}(\psi)^{-1}\,\nabla\mathcal{L}_{\text{alt}}(\psi)\right\rangle\\
&= \left\langle \mathrm{J}f(\psi)\,\nabla\mathcal{L}(\theta), \mathrm{J}f(\psi)^{-1}\, I_{x,z\,|\,\theta}(\theta)^{-1}\ \mathrm{J}f(\psi)^{-\top}\,\mathrm{J}f(\psi)\,\nabla\mathcal{L}(\theta)\right\rangle\\
&= \left\langle \nabla\mathcal{L}(\theta), I_{x,z\,|\,\theta}(\theta)^{-1}\,\nabla\mathcal{L}(\theta)\right\rangle \quad = \|\nabla\mathcal{L}(\theta)\|^2_{I_{x,z\,|\,\theta}(\theta)^{-1}}.
\end{aligned}$$

D.3 Generalized EM schemes (Theorems 1 and 2)

Theorem 1. *Under assumptions **A1**–**A3**, if the M-steps are solved up to c-multiplicative error (**A4**),*

$$\min_{t\leq T}\ \mathbb{E}[D_{A^*}(s(\theta_t),\mu_t)] \leq \frac{1}{c}\frac{\mathcal{L}(\theta_1)-\mathcal{L}^*}{T}.$$

Proof of Theorem 1. Recall the definition of the multiplicative error in **A4**,

$$\mathbb{E}[Q_{\theta_t}(\theta_{t+1}) - Q_{\theta_t}(\theta_t^*)\,|\,\theta_t] \leq (1-c)(Q_{\theta_t}(\theta_t) - Q_{\theta_t}(\theta_t^*)).$$

By adding $Q_{\theta_t}(\theta_t^*) - Q_{\theta_t}(\theta_t)$ to both sides, we get the following guarantee,

$$\mathbb{E}[Q_{\theta_t}(\theta_{t+1}) - Q_{\theta_t}(\theta_t)\,|\,\theta_t] \leq -c(Q_{\theta_t}(\theta_t) - Q_{\theta_t}(\theta_t^*)).$$

Plugging this inequality in the decomposition of the objective function (Equation 2),

$$\mathbb{E}[\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t)\,|\,\theta_t] = \mathbb{E}[Q_{\theta_t}(\theta_{t+1}) - Q_{\theta_t}(\theta_t) + \underbrace{H_{\theta_t}(\theta_t) - H_{\theta_t}(\theta_{t+1})}_{\leq 0}\,|\,\theta_t] \leq -c\,(Q_{\theta_t}(\theta_t) - Q_{\theta_t}(\theta_t^*)).$$

Using the same development as in Proposition 2, $Q_{\theta_t}(\theta_t) - Q_{\theta_t}(\theta_t^*) = D_{A^*}(s(\theta_t),\mu_t)$, and reorganizing gives

$$D_{A^*}(s(\theta_t),\mu_t) \leq \frac{1}{c}\,\mathbb{E}[\mathcal{L}(\theta_t) - \mathcal{L}(\theta_{t+1})\,|\,\theta_t].$$

Taking full expectation, averaging over all iterations and bounding $\mathbb{E}[\mathcal{L}(\theta_T)] > \mathcal{L}^*$ finishes the proof. $\square$

Theorem 2. *Under assumptions **A1**–**A3**, if the M-step at step t is solved up to ϵ_t-additive error (**A5**),*

$$\min_{t\leq T} \mathbb{E}[D_{A^*}(s(\theta_t), \mu_t)] \leq \frac{\mathcal{L}(\theta_1) - \mathcal{L}^*}{T} + \frac{1}{T}\sum_{t=1}^{T} \epsilon_t.$$

Proof of Theorem 2. Recall the definition of the additive error in **A5**,

$$\mathbb{E}[Q_{\theta_t}(\theta_{t+1}) - Q_{\theta_t}(\theta_t^*) \,|\, \theta_t] \leq \epsilon_t.$$

Plugging this inequality in the decomposition of the objective function (Equation 2),

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) \,|\, \theta_t] &= \mathbb{E}[Q_{\theta_t}(\theta_{t+1}) - Q_{\theta_t}(\theta_t) + \underbrace{H_{\theta_t}(\theta_t) - H_{\theta_t}(\theta_{t+1})}_{\leq 0} \,|\, \theta_t],\\ &\leq \mathbb{E}[Q_{\theta_t}(\theta_{t+1}) - Q_{\theta_t}(\theta_t) \,|\, \theta_t],\\ &= \mathbb{E}[Q_{\theta_t}(\theta_{t+1}) - Q_{\theta_t}(\theta_t^*) \,|\, \theta_t] + Q_{\theta_t}(\theta_t^*) - Q_{\theta_t}(\theta_t) \leq \epsilon_t + Q_{\theta_t}(\theta_t^*) - Q_{\theta_t}(\theta_t).\end{aligned}$$

Using the same developments as Proposition 2 and Corollary 2, we have $Q_{\theta_t}(\theta_t) - Q_{\theta_t}(\theta_t^*) = D_{A^*}(s(\theta_t), \mu_t)$, and

$$D_{A^*}(s(\theta_t), \mu_t) \leq \mathbb{E}[\mathcal{L}(\theta_t) - \mathcal{L}(\theta_{t+1}) \,|\, \theta_t] + \epsilon_t.$$

Taking full expectations and averaging over all iterations and bounding $\mathbb{E}[\mathcal{L}(\theta_T)] > \mathcal{L}^*$ finishes the proof. □

D.4 Relative strong-convexity and the ratio of missing information

Proposition 3. *For exponential families, the EM objective is α-strongly convex relative to A on a region Θ iff the missing information M (Equation 12) satisfies*

$$\lambda_{\max}(M(\theta)) \leq (1-\alpha) \qquad \textit{for all } \theta \in \Theta.$$

Proof of Proposition 3. That the objective is α-strong convexity relative to A is equivalent to

$$\nabla^2 \mathcal{L}(\theta) \succeq \alpha \nabla^2 A(\theta).$$

By the decomposition of the objective (Equation 2),

$$\nabla^2 \mathcal{L}(\theta) = \nabla^2 Q_\theta(\theta) - \nabla^2 H_\theta(\theta).$$

If the complete-data distribution is in the exponential family, the Hessian of the surrogate is

$$\begin{aligned}\nabla^2 Q_\theta(\theta) &= \int -\nabla^2 \log p(x, z \,|\, \theta)\, p(z \,|\, x, \theta)\, \mathrm{d}z\\ &= \int \nabla^2[A(\theta) - \langle S(x,z), \theta\rangle]\, p(z \,|\, x, \theta)\, \mathrm{d}z = \nabla^2 A(\theta),\end{aligned}$$

where $\nabla^2 A(\theta)$ is the FIM of the complete-data distribution, $I_{x,z}(\theta)$. The Hessian of the entropy term is the Fisher of the conditional distribution, $I_{z\,|\,x}(\theta)$,

$$\nabla^2 H_\theta(\theta) = \int -\nabla^2 p(z \,|\, x, \theta)\, p(z \,|\, x, \theta)\, \mathrm{d}z = I_{z\,|\,x}(\theta).$$

This gives that the relative α-strong convexity of $\mathcal{L}$ is equivalent to

$$I_{x,z}(\theta) - I_{z\,|\,x}(\theta) \succeq \alpha I_{x,z}(\theta).$$

Multiplying by the inverse of $I_{x,z}(\theta)$, which always exist if the exponential family is minimal (**A1**),

$$I - I_{x,z}(\theta)^{-1} I_{z\,|\,x}(\theta) \succeq \alpha I \qquad \iff \qquad (1-\alpha) I \succeq I_{x,z}(\theta)^{-1} I_{z\,|\,x}(\theta) = M(\theta),$$

where I is the identity matrix. This gives that $\mathcal{L}$ is α-strongly convex relative to A on a subset Θ if and only if the largest eigenvalue of the missing information is bounded by $1-\alpha$. □

D.5 Local convergence of EM (Corollaries 1 and 3)

We now present proofs for the locally convex and relatively strongly-convex settings in Corollaries 1 and 3, restated below for convenience.

Corollary 1. *For exponential families, if EM is initialized in a locally-convex region with minimum* θ^**,*

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) \leq \frac{1}{T}\,\mathrm{KL}[p(x, z \,|\, \theta_1) \| p(x, z \,|\, \theta^*)]. \tag{8}$$

Corollary 3. *Under* ***A1–A3****, if EM is initialized in a locally convex region* Θ *with minimum* $\mathcal{L}^*$ *and the ratio of missing information is bounded,* $\lambda_{\max}(M(\theta)) \leq r$*,*

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}^* \leq r(\mathcal{L}(\theta_t) - \mathcal{L}^*).$$

Both corollaries are direct consequences of Theorem 3.1 in Lu et al. (2018) with $L = 1$ if initialized in a convex or relatively $(1 - r)$-strongly convex region. We present here an alternative proof.

Theorem 3 (Simplified version of Theorem 3.1 (Lu et al., 2018)). *Let* ***A1–A3*** *hold and* $\mathcal{L}$ *be a convex and* 1*-smooth function relative to* A*, with minimum at* θ^**. Mirror descent with step-size* $\gamma = 1$*, leading to the update satisfying* $\nabla A(\theta_{t+1}) = \nabla A(\theta_t) - \nabla \mathcal{L}(\theta_t)$*, converges at the rate*

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) \leq \frac{1}{T} D_A(\theta^*, \theta_1).$$

If, in addition, $\mathcal{L}$ *is* α*-strongly convex relative to* A*, then*

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) \leq (1 - \alpha)^T D_A(\theta^*, \theta_1).$$

Proof. Recall that by definition of the update, $\nabla A(\theta_{t+1}) = \nabla A(\theta_t) - \nabla \mathcal{L}(\theta_t)$. By relative smoothness, we have

$$\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta_t \rangle + D_A(\theta_{t+1}, \theta_t).$$

We first show that the algorithm makes progress at each step, $\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t)$, by showing that

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) \leq \langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta_t \rangle + D_A(\theta_{t+1}, \theta_t) \leq -D_A(\theta_t, \theta_{t+1}).$$

Substituting the gradient by $\nabla A(\theta_t) - \nabla A(\theta_{t+1})$ we have that

$$\langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta_t \rangle + D_A(\theta_{t+1}, \theta_t) = \langle \nabla A(\theta_t) - \nabla A(\theta_{t+1}), \theta_{t+1} - \theta_t \rangle + D_A(\theta_{t+1}, \theta_t).$$

Expanding the Bregman divergence as $D_A(\theta_{t+1}, \theta_t) = A(\theta_{t+1}) - A(\theta_t) - \langle \nabla A(\theta_t), \theta_{t+1} - \theta_t \rangle$, we get the simplification

$$\begin{aligned}
&= \langle \nabla A(\theta_t) - \nabla A(\theta_{t+1}), \theta_{t+1} - \theta_t \rangle + A(\theta_{t+1}) - A(\theta_t) - \langle \nabla A(\theta_t), \theta_{t+1} - \theta_t \rangle, \\
&= -\langle \nabla A(\theta_{t+1}), \theta_{t+1} - \theta_t \rangle + A(\theta_{t+1}) - A(\theta_t) \\
&= -D_A(\theta_t, \theta_{t+1}) \leq 0.
\end{aligned}$$

We now relate the progress to the Bregman divergence to the minimum. We will show that

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta^*) \leq D_A(\theta^*, \theta_t) - D_A(\theta^*, \theta_{t+1}).$$

Starting from relative smoothness,

$$\begin{aligned}
\mathcal{L}(\theta_{t+1}) &\leq \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta_t \rangle + D_A(\theta_{t+1}, \theta_t), \\
&= \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta^* + \theta^* - \theta_t \rangle + D_A(\theta_{t+1}, \theta_t), && (\pm \langle \nabla \mathcal{L}(\theta_t), \theta^* \rangle) \\
&= \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta^* - \theta_t \rangle + \langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta^* \rangle + D_A(\theta_{t+1}, \theta_t).
\end{aligned}$$

By convexity, we have that $\mathcal{L}(\theta^*) \geq \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta^* - \theta_t \rangle$ and

$$\leq \mathcal{L}(\theta^*) + \langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta^* \rangle + D_A(\theta_{t+1}, \theta_t).$$

Using that the update satisfies $\nabla A(\theta_{t+1}) = \nabla A(\theta_t) - \nabla \mathcal{L}(\theta_t)$, we can rewrite the gradient as

$$\begin{aligned}
&= \mathcal{L}(\theta^*) + \langle \nabla A(\theta_t) - \nabla A(\theta_{t+1}), \theta_{t+1} - \theta^* \rangle + D_A(\theta_{t+1}, \theta_t), \\
&= \mathcal{L}(\theta^*) + \langle \theta^* - \theta_{t+1}, \nabla A(\theta_{t+1}) - \nabla A(\theta_t) \rangle + D_A(\theta_{t+1}, \theta_t).
\end{aligned}$$

Using the three point property, $D_A(\theta^*, \theta_t) = D_A(\theta^*, \theta_{t+1}) + \langle \theta^* - \theta_{t+1}, \nabla A(\theta_{t+1}) - \nabla A(\theta_t)\rangle + D_A(\theta_{t+1}, \theta_t)$ and

$$= \mathcal{L}(\theta^*) + D_A(\theta^*, \theta_t) - D_A(\theta^*, \theta_{t+1}).$$

Reorganizing the terms yields the inequality $\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta^*) \leq D_A(\theta^*, \theta_t) - D_A(\theta^*, \theta_{t+1})$.

Using that the algorithm makes progress and summing all iterations yields

$$T(\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*)) \leq \sum_{t=1}^{T} \mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta^*) \leq \sum_{t=1}^{T} D_A(\theta^*, \theta_t) - D_A(\theta^*, \theta_{t+1}) \leq D_A(\theta^*, \theta_1).$$

Dividing by T finishes the proof for the convex case.

For the relatively strongly-convex case, we will show that the Bregman divergence also converges linearly,

$$D_A(\theta^*, \theta_{t+1}) \leq (1-\alpha)^t D_A(\theta^*, \theta_1).$$

Combining this contraction with earlier result that $\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta^*) \leq D_A(\theta^*, \theta_t) - D_A(\theta^*, \theta_{t+1})$ implies

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta^*) \leq D_A(\theta^*, \theta_t) - D_A(\theta^*, \theta_{t+1}) \leq D_A(\theta^*, \theta_t) \leq (1-\alpha)^t D_A(\theta^*, \theta_1).$$

In addition to the three point property, we will use the following two results to show the linear rate of convergence. By the relative α-strong convexity of $\mathcal{L}$,

$$\mathcal{L}(\theta^*) \geq \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta^* - \theta_t\rangle + \alpha D_A(\theta^*, \theta_t) \implies \quad \langle \nabla \mathcal{L}(\theta_t), \theta^* - \theta_t\rangle \leq \mathcal{L}(\theta^*) - \mathcal{L}(\theta_t) - \alpha D_A(\theta^*, \theta_t), \quad \text{(A)}$$

And by the first result we showed, the algorithm makes progress proportional to $D_A(\theta_t, \theta_{t+1})$,

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) \leq -D_A(\theta_t, \theta_{t+1}) \implies \qquad D_A(\theta_t, \theta_{t+1}) \leq \mathcal{L}(\theta_t) - \mathcal{L}(\theta_{t+1}) \quad \text{(B)}$$

Using the three point property, we can expand the divergence as

$$D_A(\theta^*, \theta_{t+1}) = D_A(\theta^*, \theta_t) + \langle \theta^* - \theta_t, \nabla A(\theta_t) - \nabla A(\theta_{t+1})\rangle + D_A(\theta_t, \theta_{t+1}).$$

Replacing $\nabla A(\theta_t) - \nabla A(\theta_{t+1})$ by the gradient at θ_t, we have

$$= D_A(\theta^*, \theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta^* - \theta_t\rangle + D_A(\theta_t, \theta_{t+1}).$$

Using the relative α-strong convexity of $\mathcal{L}$ (A),

$$\leq (1-\alpha) D_A(\theta^*, \theta_t) + (\mathcal{L}(\theta^*) - \mathcal{L}(\theta_t)) + D_A(\theta_t, \theta_{t+1}).$$

Using the progress bound (B),

$$\begin{aligned} &\leq (1-\alpha) D_A(\theta^*, \theta_t) + (\mathcal{L}(\theta^*) - \mathcal{L}(\theta_t)) + (\mathcal{L}(\theta_t) - \mathcal{L}(\theta_{t+1})), \\ &= (1-\alpha) D_A(\theta^*, \theta_t) + (\mathcal{L}(\theta^*) - \mathcal{L}(\theta_{t+1})). \end{aligned}$$

As $\mathcal{L}(\theta^*) \leq \mathcal{L}(\theta_{t+1})$, we get that $D_A(\theta^*, \theta_{t+1}) \leq (1-\alpha) D_A(\theta^*, \theta_t)$. Recursing finishes the proof,

$$D_A(\theta^*, \theta_{t+1}) \leq (1-\alpha)^t D_A(\theta^*, \theta_1).$$

□

E Supplementary material for Section 5.4: EM for General Models

This sections extends the results on stationarity in Section 5 to handle cases where the objective and the surrogate can be non-differentiable. A simple example of this setting is a mixture of Laplace distributions. It is still possible to optimize the M-step, but the theory does not apply as the Laplace is not in the exponential family. The main problem for the analysis of non-differentiable, non-convex objectives is that the progress at each step need not be related to the gradient (if it is even defined at the current point). Asymptotic convergence can still be shown (Chrétien and Hero, 2000; Tseng, 2004), but non-asymptotic results are not available without stronger assumptions, such as the Kurdyka-Łojasiewicz inequality or weak convexity.

Instead of focusing on the progress of the M-step, we look here at the progress of the E-step under the assumption that the conditional distribution over the latent variables $p(z \,|\, x, \theta)$ is in the exponential family. This is a strictly weaker assumption, as it is implied if the complete-data distribution $p(x, z \,|\, \theta)$ is an exponential family distribution, but holds more generally. For example, it is satisfied by any finite mixture, even if the mixture components are non-differentiable, as for the mixture of Laplace distributions. As a tradeoff, however, the resulting convergence results only describe the stationarity of the parameters controlling the latent variables.

To analyse the E-step, we use the formulation of EM as a block-coordinate optimization problem. Let $q(z \,|\, \phi)$ be an exponential family distribution in the same family as $p(z \,|\, x, \theta)$, such that $\min_\phi \mathrm{KL}[q(z \,|\, \phi) \| p(z \,|\, x, \theta)] = 0$. We can write the E-step and M-step as an alternating optimization procedure on the augmented objective $\mathcal{L}^+$,

$$\mathcal{L}^+(\theta, \phi) = -\int \log\left(\frac{p(x, z \,|\, \theta)}{q(z \,|\, \phi)}\right) q(z \,|\, \phi)\,\mathrm{d}z \qquad \text{such that} \qquad \mathcal{L}(\theta) = \min_\phi \mathcal{L}^+(\theta, \phi),$$

The parameters ϕ and θ need not be defined on the same space, as ϕ only controls the conditional distribution over the latent variables and θ controls the complete-data distribution. The E and M steps then correspond to

$$\text{E-step:} \qquad \phi_{t+1} = \arg\min_\phi \mathcal{L}^+(\theta_t, \phi), \qquad \text{M-step:} \qquad \theta_{t+1} = \arg\min_\theta \mathcal{L}^+(\theta, \phi_{t+1}).$$

Two gradients now describe stationarity; the gradient of the M-step, $\nabla_\theta \mathcal{L}^+(\theta_t, \phi_{t+1})$, which we studied before, and the gradient of the E-step, $\nabla_\phi \mathcal{L}^+(\theta_t, \phi_t)$. Let S and A be the sufficient statistics and log-partition function of $q(z \,|\, \phi)$, and the natural and equivalent mean parameters be denoted by (ϕ_t, μ_t). We show the following, which is the analog of Corollary 2 for the conditional distribution over the latent variables $q(z \,|\, \phi)$.

Theorem 4. *Let Assumption* ***A2*** *and* ***A3*** *hold, and let θ be the parameters of the complete-data distribution $p(x, z \,|\, \theta)$. If the conditional distribution over the latent variables $q(z \,|\, \phi)$ is a minimal exponential family distribution with natural and mean parameters (ϕ, μ),*

$$\min_{t \leq T} D_A(\phi_{t+1}, \phi_t) \leq \frac{\mathcal{L}(\theta_1) - \mathcal{L}^*}{T}.$$

This implies convergence of the gradient in KL divergence as the (natural) gradient is $\nabla_\mu \mathcal{L}^+(\theta_t, \phi_t) = \phi_t - \phi_{t+1}$.

Proof of Theorem 4. Let us start by bounding the progress on the overall objective by the progress of the E-step;

$$\begin{aligned}\mathcal{L}(\theta_t) - \mathcal{L}(\theta_{t-1}) &= \overbrace{\mathcal{L}^+(\theta_t, \phi_{t+1}) - \mathcal{L}^+(\theta_{t-1}, \phi_t)}^{\text{Progress of the joint EM step}}, \\ &= \underbrace{\mathcal{L}^+(\theta_t, \phi_{t+1}) - \mathcal{L}^+(\theta_t, \phi_t)}_{\text{Progress of the E-step}} + \underbrace{\mathcal{L}^+(\theta_t, \phi_t) - \mathcal{L}^+(\theta_{t-1}, \phi_t)}_{\text{Progress of the M-step}} \leq \mathcal{L}^+(\theta_t, \phi_{t+1}) - \mathcal{L}^+(\theta_t, \phi_t).\end{aligned}$$

The last inequality holds as the M-step is guarantee to make progress. To show that the progress of the E-step is the KL divergence between $q(z \,|\, \phi)$ and $p(z \,|\, x, \theta_t)$ we use the following substitution,

$$\begin{aligned}\mathcal{L}^+(\theta, \phi) = -\int \log \frac{p(x, z \,|\, \theta)}{q(z \,|\, \phi)} q(z \,|\, \phi)\,\mathrm{d}z &= -\int \log \frac{p(z \,|\, x, \theta)}{q(z \,|\, \phi)} q(z \,|\, \phi)\,\mathrm{d}z - \int \log p(x \,|\, \theta) q(z \,|\, \phi)\,\mathrm{d}z \\ &= \mathrm{KL}[q(z \,|\, \phi) \| p(z \,|\, x, \theta)] - \log p(x \,|\, \theta).\end{aligned}$$

Plugging the substitution in $\mathcal{L}^+(\theta_t, \phi_{t+1}) - \mathcal{L}^+(\theta_t, \phi_t)$ yields

$$\mathcal{L}^+(\theta_t, \phi_{t+1}) - \mathcal{L}^+(\theta_t, \phi_t) = \underbrace{\mathrm{KL}[q(z \,|\, \phi_{t+1}) \| p(z \,|\, x, \theta_t)]}_{=0} - \mathrm{KL}[q(z \,|\, \phi_t) \| p(z \,|\, x, \theta_t)],$$

where the first term is 0 if $p(z \mid x, \theta)$ and $q(z \mid \phi)$ are in the same exponential family and ϕ_{t+1} is the exact solution

$$s(\theta_t) = \mathbb{E}_{p(z \mid x, \theta_t)}[S(z)], \qquad \phi_{t+1} = \nabla A^*(s(\theta_t)).$$

Combining the bounds so far, we have that

$$\mathcal{L}(\theta_t) - \mathcal{L}(\theta_{t-1}) \leq \mathcal{L}^+(\theta_t, \phi_{t+1}) - \mathcal{L}^+(\theta_t, \phi_t) \leq -\mathrm{KL}[q(z \mid \phi_t) \| p(z \mid x, \theta_t)].$$

To relate the progress to the gradient, we express the KL divergence as a Bregman divergence in mean parameters,

$$\mathrm{KL}[q(z \mid \phi_t) \| p(z \mid x, \theta_t)] = D_{A^*}(\mu_t, s(\theta_t)).$$

The gradient with respect to the mean parameters at (ϕ_t, μ_t) is then

$$\begin{aligned}
\nabla_\mu \mathcal{L}^+(\theta_t, \nabla A^*(\mu)) \,|_{\,\mu=\mu_t} &= \nabla_\mu \mathrm{KL}[q(z \mid \nabla A^*(\mu)) \| p(z \mid x, \theta_t)] \,|_{\,\mu=\mu_t}, \\
&= \nabla_\mu D_{A^*}(\mu, s(\theta_t)) \,|_{\,\mu=\mu_t}, \\
&= \nabla_\mu [A^*(\mu) - A^*(s(\theta_t)) - \langle \nabla A^*(s(\theta_t)), \mu - s(\theta_t) \rangle\,] \,|_{\,\mu=\mu_t}, \\
&= \nabla A^*(\mu_t) - \nabla A^*(s(\theta_t)) = \phi_t - \phi_{t+1}.
\end{aligned}$$

We can then express the update of the E-step from ϕ_t to ϕ_{t+1} as a mirror descent step, updating the natural parameters using the gradient with respect to the natural parameters,

$$\phi_{t+1} = \phi_t - \nabla_\mu \mathcal{L}^+(\theta_t, \nabla A^*(\mu_t)).$$

To express this update in natural parameters only, recall from Appendices A.2 and A.4 that $\nabla^2 A^*(\mu_t) = [\nabla^2 A(\phi_t)]^{-1}$ and that $\nabla^2 A(\phi_t)$ is the Fisher information matrix of the distribution $q(z \mid \phi)$, $I_{z \mid \phi}(\phi_t)$. The update is then equivalent to a natural gradient update in natural parameters, as

$$\nabla_\mu \mathcal{L}^+(\theta_t, \nabla A^*(\mu_t)) = \nabla^2 A^*(\mu_t)\, \nabla_\phi \mathcal{L}^+(\theta_t, \phi_t) = [I_{z \mid \phi}(\phi_t)]^{-1}\, \nabla_\phi \mathcal{L}^+(\theta_t, \phi_t).$$

Using those expression for the KL divergence and parameter updates yields the bound

$$\mathcal{L}(\theta_t) - \mathcal{L}(\theta_{t-1}) \leq -\mathrm{KL}[q(z \mid \phi_t) \| p(z \mid x, \theta_t)] = -D_A(\phi_t - \nabla_\mu \mathcal{L}^+(\theta_{t+1}, \nabla A^*(\mu_t)), \phi_t).$$

Reorganizing terms gives

$$D_A(\phi_t - \nabla_\mu \mathcal{L}^+(\theta_t, \nabla A^*(\mu_t)), \phi_t) \leq \mathcal{L}(\theta_{t-1}) - \mathcal{L}(\theta_t).$$

And averaging over all iterations and bounding $\mathcal{L}(\theta_T) > \mathcal{L}^*$ finishes the proof. □