Convergence rate of MAP estimates for the exponential family

Rémi Le Priol

October 2020

1 Background

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

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

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

with log-partition function

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

Recall that A verifies the two following identities

$$\nabla A(\theta) = \mathbb{E}_{p(X|\theta)} [T(X)] =: \mu \tag{3}$$

$$\nabla^2 A(\theta) = \mathrm{Cov}_\theta[T(X)] > 0 \tag{4}$$

where μ is called the mean parameter. The second identity entails that A is strictly-convex. The conjugate prior for this distribution is

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

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

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

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

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

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

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

Minimizing this expression over θ yields the Maximum A Posteriori estimate

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

such that the MAP is

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

The MAP estimate is a random quantity.

2 Straightforward Convergence Rate

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

$$\begin{aligned}
&\mathbb{E}_{x \sim p(.|\theta^*)} \left[-\log p(x|\theta) + \log p(x|\theta^*) \right] && (10)\\
=&D_{\mathrm{KL}}(p(.|\theta^*)||p(.|\theta)) && (11)\\
=&\mathcal{B}_A(\theta||\theta^*) && (12)\\
=&\mathcal{B}_{A^*}(\mu^*||\mu) && (13)\\
=&A^*(\mu^*) + A(\theta) - \langle \theta, \mu^* \rangle && (14)
\end{aligned}$$

where A^* is the entropy, the convex conjugate of the log-partition. The relationship between Bregman divergences and Fenchel conjugacy is well explained in Wainwright and Jordan (2008), and Agarwal and Daumé (2010). The question is: how does this quantity behave when θ is the maximum-likelihood or the MAP estimate ? Can we get bounds – in expectation or high-probability ?

$$\mathcal{B}_{A^*}\left(\mathbb{E}[T(X)]; \frac{1}{n} \textstyle\sum_i T(x_i) \right) \tag{15}$$

$$\mathcal{B}_{A^*}\left(\mathbb{E}[T(X)]; \frac{n_0\mu_0 + \sum_i T(x_i)}{n_0 + n} \right) \tag{16}$$

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

$$\mathcal{B}_{A^*}(\mu^*||\mu) \leq L\|\mu^* - \mu\| + \|\theta\|\|\mu^* - \mu\| \leq 2L\|\mu^* - \mu\| \tag{17}$$

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

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

$$\mathcal{B}_{A^*}(\mu^*||\mu) \leq \frac{L}{2}\|\mu^* - \mu\|^2 \tag{18}$$

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. If $\operatorname{Var} T(X) = \sigma^2$, then

$$\mathbb{E}\left\|\mu^* - \frac{1}{n}\textstyle\sum_i T(x_i)\right\|^2 = \frac{\sigma^2}{n} \tag{19}$$

eg the variance of the mean is n times smaller than the variance of the samples. We can extend this result to a Manahalobis distance with any matrix $\boldsymbol{M}$

$$\mathbb{E}\left\|\mu^* - \frac{1}{n}\textstyle\sum_i T(x_i)\right\|_{\boldsymbol{M}}^2 = \frac{1}{n}\operatorname{Tr}(\boldsymbol{M\Sigma}) \tag{20}$$

where $\boldsymbol{\Sigma}$ is the covariance of $T(X)$. This bound is equal to σ^2/n, the variance divided by the number of samples, when the metric is the identity $\boldsymbol{M} = \boldsymbol{I}$. Adding a reference mean μ_0 to get the MAP,

$$\begin{aligned}\mathbb{E}\left\|\mu^* - \frac{n_0\mu_0 + \sum_i T(x_i)}{n_0 + n}\right\|^2 &= \frac{n}{(n+n_0)^2}\sigma^2 + \frac{n_0^2}{(n+n_0)^2}\|\mu^* - \mu_0\|^2 &(21)\\ &= O\left(\frac{\sigma^2}{n}\right) + O\left(\frac{\|\mu^* - \mu_0\|^2}{n^2}\right) &(22)\end{aligned}$$

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 !

3 Stochastic Bregman Proximal Point

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

$$\begin{aligned}\hat{\theta}_{n+1} &= \operatorname*{argmin}_\theta -\log p(x_{n+1}|\theta) + (n_0+n)\mathcal{B}_A(\theta||\hat{\theta}_n) &(23)\\ &= \operatorname*{argmin}_\theta f(\theta; x_{n+1}) + \frac{1}{\gamma_n}\mathcal{B}_A(\theta||\hat{\theta}_n)\ . &(24)\end{aligned}$$

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

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

4 Stochastic Mirror Descent (SMD)

4.1 MAP as SMD

Let $f(\theta) := \mathbb{E}_x[-\log p(x|\theta)] = -\langle \mu_*, \theta\rangle + A(\theta)$. In words, f is linear modification of a convex function A, which we can access only through noisy estimates of μ_*. It turns out that the MAP estimate can also be seen as the iterates of stochastic mirror descent with mirror map A. First let's recall mirror descent iteration

$$\theta_{t+1} := \operatorname*{argmin}_\theta \gamma \ell_f(\theta; \theta_t) + \mathcal{B}_A(\theta||\theta_t) \tag{25}$$
$$= \nabla A^*(\nabla A(\theta_t) - \gamma \nabla f(\theta_t)) \tag{26}$$

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. Indeed plugging in definitions of f and μ yields

$$\mu_{t+1} = \mu_t - \gamma(\mu_t - \mu_*) \tag{27}$$
$$\implies \mu_t = \mu_* + (1-\gamma)^t(\mu_0 - \mu_*) \tag{28}$$

which shows exponential convergence and 1-step convergence when $\gamma = 1$. Back to our sheep, the MAP iteration can be cast as stochastic mirror descent (SMD), with $g_t = \nabla f(\theta_t, x_{t+1})$ a stochastic estimate of $\nabla f(\theta_t)$

$$\hat{\theta}_{n+1} = \operatorname*{argmin}_\theta -\langle T(x_{t+1}), \theta\rangle + A(\theta) + (n_0 + n)\mathcal{B}_A(\theta||\theta_n) \tag{29}$$
$$= \operatorname*{argmin}_\theta -\langle T(x_{t+1}), \theta\rangle + A(\theta_n) + \langle \nabla A(\theta_n), \theta - \theta_n\rangle + (n_0 + n + 1)\mathcal{B}_A(\theta||\theta_n) \tag{30}$$
$$= \operatorname*{argmin}_\theta \ell_f(\theta; \theta_n, x_{t+1}) + (n_0 + n + 1)\mathcal{B}_A(\theta||\theta_n) \tag{31}$$

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 $\gamma_n = \frac{1}{n_0+n+1}$.

4.2 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. f is μ

strongly-convex and L-smooth relative to h if

$$
\begin{aligned}
& f - \mu h \;\text{ convex and } Lh - f \text{ convex} && (32)\\
\iff & f(x) + \langle \nabla f(x), y - x\rangle + \mu \mathcal{B}_h(y||x) \leq f(y) && (33)\\
& \text{and} \quad f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + L\mathcal{B}_h(y||x), \forall x, y && (34)\\
\iff & \mu \mathcal{B}_h(y||x) \leq \mathcal{B}_f(y||x) \leq L\mathcal{B}_h(y||x), \forall x, y && (35)\\
\iff & \mu \nabla^2 h(x) \leq \nabla^2 f(x) \leq L \nabla^2 h(x), \forall x && (36)
\end{aligned}
$$

where the last equivalence holds only when f and h are twice differentiable. This sweet generalization of smoothness and strong-convexity transfers the Loewner partial order 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

$$x_{t+1} = \operatorname*{argmin}_x \langle \nabla f(x_t), x\rangle + \mathcal{B}_h(x||x_t) \tag{37}$$

is upper bounded by the simple formula

$$\implies f(x_t) - f(x_*) \leq \frac{L\mathcal{B}_h(x_*||x_0)}{t} \,. \tag{38}$$

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.

4.3 Relative smoothness for SMD

Now our setting is Stochastic Mirror Descent (SMD), meaning at each step we observe a random unbiased estimate gradient. This setting was studied by Hanzely and Richtárik (2018), who proved in the smooth strongly-convex case with tail averaging : with constant step-size, linear convergence down to a variance ball, and with step-size $\gamma_t = n_0 + t$ a rate $\tilde{O}(\frac{1}{t})$. These results match the rates for standard SGD.

This is very interesting for us, but the variance hyper-parameter is oddly defined. Let g_t be the random gradient at step t (coming from data point x_t), and θ_{t+1} the next iterate. Then the variance bound σ^2 is an upper bound on the covariance between the gradient update $-g_t$ and the descent direction $\theta_{t+1} - \theta_t = \nabla A^*(\nabla A(\theta_t) - \gamma_t g_t) - \theta_t$ that should hold for all time steps

$$
\begin{aligned}
& \mathrm{Cov}(-g_t, \theta_{t+1} - \theta_t | X_{1\dots t}) \leq \gamma_t \sigma^2 && (39)\\
& = \mathbb{E}\left[\langle -g_t, \theta_{t+1}\rangle\right] - \langle \mathbb{E}[-g_t], \mathbb{E}[\theta_{t+1}]\rangle && (40)\\
& = \mathbb{E}\left[\langle \nabla f(\theta_t) - g_t, \theta_{t+1}\rangle\right] && (41)
\end{aligned}
$$

where γ_t is the step-size and expectations are conditional on the past. Remark that when we plug in $A^* = \|.\|^2$, we recover the gradient variance typical of SGD. In our case,

$$\nabla f(\theta_t) - g_t = T(X_{t+1}) - \mathbb{E}[T(X)] \tag{42}$$

so that σ is really a bound on the covariance of the sufficient statistics with another variable

$$\begin{aligned}
&\mathrm{Cov}(T(X_{t+1}), \theta_{t+1} | X_{1\dots t}) \leq \gamma_t \sigma^2 && (43)\\
=\,&\mathrm{Cov}(T(X_{t+1}), \nabla A^* \left(\gamma_t T(X_{t+1}) + (1-\gamma_t)\mu_t\right) | X_{1\dots t})\ . && (44)
\end{aligned}$$

In other words, we need for all $\gamma \in (0,1)$, and all μ in the interior of the marginal polytope,

$$\mathrm{Cov}_X \Big(T(X), \nabla A^*(\gamma T(X) + (1-\gamma)\mu) \Big) \leq \gamma \sigma^2\ . \tag{45}$$

When we plug this bound into the expected suboptimality formula (15) with the maximum likelihood estimate $\hat{\mu} = \frac{1}{n}\sum_i T(X_i)$, assuming all $T(x)$ belong to the marginal polytope, we get

$$\begin{aligned}
\mathbb{E}_{X_{1\dots n}}\left[\mathcal{B}_{A^*}(\mu_* || \hat{\mu})\right] &= A^*(\mu_*) \overbrace{- \mathbb{E}[A^*(\hat{\mu})]}^{\leq -A^*(\mathbb{E}[\hat{\mu}])} \\
&\quad + \frac{1}{n}\sum_i \mathbb{E}[\underbrace{\mathbb{E}[\langle T(X_i) - \mu_*; \nabla A^*(\hat{\mu})\rangle | X_j, j \neq i]}_{\leq \sigma^2/n (45)}] \leq \frac{\sigma^2}{n} \qquad (46)
\end{aligned}$$

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 (45). In words, this variance assumption on $T(X)$ and A^* immediately gives us a bound on the suboptimality. We can also apply this result to the MAP estimate $\hat{\mu} = \frac{n_0\mu_0 + \sum_i T(X_i)}{n_0 + n}$ but I did not manage to reach a satisfying conclusion about the bias. Note $\tilde{\mu} = \mathbb{E}[\hat{\mu}] = \frac{n_0\mu_0 + n\mu_*}{n_0+n}$ and $\gamma_0 = \frac{n_0}{n_0+n}$ and the step-size $\gamma = \frac{1}{n_0+n}$.

$$\begin{aligned}
\mathbb{E}_{X_{1\dots n}}\left[\mathcal{B}_{A^*}(\mu_* || \hat{\mu})\right] &\leq A^*(\mu_*) - A^*(\tilde{\mu}) + \gamma_0 \langle \mu_0 - \mu_*; \mathbb{E}[\hat{\theta}]\rangle + (1-\gamma_0)\gamma\sigma^2 && (47)\\
&= \mathcal{B}_{A^*}(\mu_*; \tilde{\mu}) - \langle \tilde{\mu} - \mu_*; \tilde{\theta} - \mathbb{E}[\hat{\theta}]\rangle + \frac{n}{(n_0+n)^2}\sigma^2 && (48)
\end{aligned}$$

so we recover the $O(n^{-1})$ for the variance and we are still looking for $O(n^{-2})$ rate for the bias. The Bregman term in the bias should be in this spirit, given a quadratic approximation

$$\mathcal{B}_{A^*}(\mu_*; \tilde{\mu}) = \mathcal{B}_{A^*}(\mu_*; \mu_* + \gamma_0(\mu_0 - \mu_*)) \approx \frac{1}{2}\gamma_0^2 \|\mu_0 - \mu_*\|^2_{\mathbf{\Sigma}_*^{-1}} \tag{49}$$

where $\mathbf{\Sigma}_* = \mathrm{Cov}_{\mu_*}(T(X))$. RLP:beware of the order, this might be inaccurate However the scalar product is harder to bound. Hopefully it could be positive.

If not we know that $\tilde{\mu} - \mu_* = \frac{n_0}{n_0+n}(\mu_0 - \mu_*) = O(n^{-1})$, and we can hope that the same holds for $\tilde{\theta} - \mathbb{E}[\hat{\theta}] = \nabla A^*(\mathbb{E}[\hat{\mu}]) - \mathbb{E}[\nabla A^*(\hat{\mu})]$, which kinda measures the non-linearity, or the curvature of A^*. Using a simple cauchy-schwartz, and the hessian of the conjugate, we might be able to get something.

TODO The question remains of : when is such a bound valid? For which values of σ^2. To answer such questions, we need to turn to examples.

The other article Lu (2019) derives another convergence rate for SMD under another gradient variance assumption.

Also Raskutti and Mukherjee (2015) only gives asymptotic results.

5 Self-Concordance

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

$$A(\eta, \Lambda) = \frac{1}{2}\eta^\top \Lambda^{-1} \eta - \log\det(\Lambda) \tag{50}$$

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

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

$$|f'''(x)| \leq 2f''(x)^{\frac{3}{2}} \ . \tag{51}$$

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

5.1 Suboptimality and Newton Decrement

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

$$D(x)^2 = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \ . \tag{52}$$

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

$$f(x) - f* \leq -D(x) - \log(1 - D(x)) \ . \tag{53}$$

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

$$f(x) - f^* \leq D(x)^2 = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \ . \tag{54}$$

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

$$f(\theta) = A(\theta) - A(\theta^*) - \langle \mu^*, \theta - \theta^* \rangle \tag{55}$$

$$\nabla f(\theta) = \mu - \mu^* = \mathbb{E}_{p(X|\theta)}\left[T(X)\right] - \mathbb{E}_{p(X|\theta^*)}\left[T(X)\right] \tag{56}$$

$$\nabla^2 f(\theta) = \mathbf{\Sigma}(\theta) = \mathrm{Cov}_{p(X|\theta)}[T(X)] \tag{57}$$

so that we get the bound.

$$\mathcal{B}_{A^*}(\mu^*||\mu) \leq D(\theta)^2 = \|\mu^* - \mu(\theta)\|^2_{\mathbf{\Sigma}(\theta)^{-1}} \leq 0.46 \tag{58}$$

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

$$\nabla h(\mu) = \theta - \theta^* \tag{59}$$

$$\nabla^2 h(\mu) = \nabla^2 A^*(\mu) = \nabla^2 A(\theta)^{-1} = \mathrm{Cov}_{p(X|\theta)}[T(X)]^{-1} \tag{60}$$

$$\implies D(\mu) = \mathrm{Var}_{p(X|\theta)}[(\theta - \theta^*)^T T(X)] \ . \tag{61}$$

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

References

References

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

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

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

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

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

Lu, H. (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.

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 Three point lemma

Proofs of convergence for mirror descent with relative smoothness rely on a specific property of Bregman divergences, familiar to information geometry folks. If x_+ is solution to

$$\min_{x \in C} \overbrace{f(x) + \mathcal{B}_h(x \| y)}^{\phi(x)} \tag{62}$$

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 (we will drop the index, and use a semicolon instead of double bars to lighten the notation), and y is some reference vector. Then

$$\forall x, f(x) + \mathcal{B}(x; y) \geq f(x_+) + \mathcal{B}(x; x_+) + \mathcal{B}(x_+; y)\ . \tag{63}$$

This property is an analog of the Pythagorean theorem for generalized projections (setting $f = 0$ and $h = \|.\|^2$).

Proof. The first order optimality condition is

$$\begin{aligned} &\langle \nabla \phi(x_+), x - x_+ \rangle \geq 0, \forall x & (64)\\ =&\langle \nabla f(x_+) + \nabla h(x_+) - \nabla h(y), x - x_+ \rangle & (65) \end{aligned}$$

This proof relies on another property called three point property

$$\langle \nabla h(x_+) - \nabla h(y), x - x_+ \rangle = \mathcal{B}(x; y) - \mathcal{B}(x; x_+) - \mathcal{B}(x_+; y) \tag{66}$$

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

$$f(x_+) + \langle \nabla f(x_+), x - x_+ \rangle \leq f(x)\ . \tag{67}$$

Putting it all together we get

$$\begin{aligned} 0 &\leq \langle \nabla f(x_+), x - x_+ \rangle + \langle \nabla h(x_+) - \nabla h(y), x - x_+ \rangle & (68)\\ &\leq f(x) - f(x_+) + \mathcal{B}(x; y) - \mathcal{B}(x; x_+) - \mathcal{B}(x_+; y) & (69) \end{aligned}$$

which concludes the proof. □

B Fenchel conjugate motivation to self-concordance

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

$$\nabla f \circ \nabla f^* = \mathrm{Id} \tag{70}$$

$$\nabla f^* \circ \nabla f = \mathrm{Id} \tag{71}$$

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

$$\nabla^2 f(x) \nabla^2 f^*(x^*) = I_n \tag{72}$$

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

$$f''(x) f^{*\prime\prime}(f'(x)) = 1, \forall x \tag{73}$$
$$\implies f'''(x) f^{*\prime\prime}(f'(x)) + f''(x)^2 f^{*\prime\prime\prime}(f'(x)) = 0 \tag{74}$$
$$\implies \frac{f'''(x)}{f''(x)^{\frac{3}{2}}} + \frac{f^{*\prime\prime\prime}(x^*)}{f^{*\prime\prime}(x^*)^{\frac{3}{2}}} = 0 \tag{75}$$

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

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

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

$$\nabla^2 f^{-\frac{1}{2}} \nabla^3 f \nabla^2 f^{-1} + \nabla^2 f^{*-\frac{1}{2}} \nabla^3 f^* \nabla^2 f^{*-1} = 0 \tag{76}$$

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

C MAP on Graphs

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

$$p(X|\theta) = \prod_i p(X_i | X_{G(i)}; \theta_i) \tag{77}$$

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

$$p(\theta) = \prod_i p(\theta_i) \; . \tag{78}$$

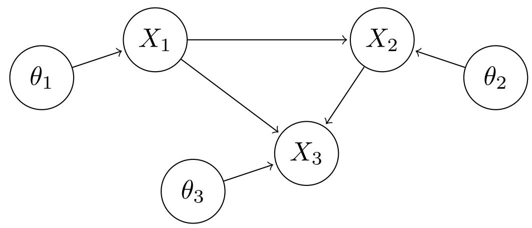

Figure 1: A graph G' factorizing (θ, X). Although the graph restricted on X does not encode any conditional independence, G' does on the joint distribution.

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

$$\begin{aligned}
p(\theta|X) &\propto p(X|\theta)p(\theta) && (79)\\
&= \prod_i p(X_i|X_{G(i)};\theta_i)p(\theta_i) && (\perp\!\!\!\perp_i \theta_i) \quad (80)\\
&= \prod_i p(X_i|X_{G(i)};\theta_i)p(\theta_i|X_{G(i)}) && (\theta_i \perp\!\!\!\perp X_{G(i)}) \quad (81)\\
&= \prod_i p(X_i, \theta_i|X_{G(i)}) && (82)\\
&= \prod_i p(\theta_i|X_i, X_{G(i)})p(X_i|X_{G(i)}) && (83)\\
\implies p(\theta|X) &= \prod_i p(\theta_i|X_i, X_{G(i)}) \; . && (84)
\end{aligned}$$

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

C.1 Equality of directions for 2 categorical variables

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

$$\underbrace{\mathrm{Dir}_{K^2}((\gamma_{i,j})_{i,j=1}^K)}_{p(\omega)} \equiv \underbrace{\mathrm{Dir}_K(\gamma\,\mathbf{1})}_{p(\mu)} \otimes \left(\bigotimes_{i=1}^K \underbrace{\mathrm{Dir}_K((\gamma_{i,j})_j)}_{p(\nu_i)} \right) \tag{85}$$

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

$$p(\mu|\mathcal{A}) = \mathrm{Dir}(\gamma\,\mathbf{1} + \sum_i A_i) \tag{86}$$

$$p(\nu_k|\mathcal{A}, \mathcal{B}) = \mathrm{Dir}(\gamma_{k,:} + \sum_i A_{i,k} B_i) \tag{87}$$

$$p(\omega|\mathcal{A}, \mathcal{B}) = \mathrm{Dir}(\gamma + \sum_i A_i B_i^\top) \tag{88}$$

where $A_i B_i^\top$ is the one hot matrix encoding of A, B. These three posteriors are obtained independently of each other following rules of calculus for Dirichlet distributions. Yet they happen to define the same distribution on distributions, as we verify below with the two equalities from equation (85).

$$\left(\gamma + \sum_i A_i B_i^\top\right)_{k,l} = \left(\gamma_{k,:} + \sum_i A_{i,k} B_i\right)_l \tag{89}$$

$$\left(\gamma + \sum_i A_i B_i^\top\right)\mathbf{1} = \gamma\,\mathbf{1} + \sum_i A_i \;. \tag{90}$$

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

$$\hat{\omega}^{\mathrm{MAP}} = \frac{\gamma + \sum_i A_i B_i^\top}{\mathbf{1}^\top(\gamma + \sum_i A_i B_i^\top)\,\mathbf{1}} = \frac{\gamma + \sum_i A_i B_i^\top}{n_0 + n} \tag{91}$$

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

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