1 Basic Properties of Fixed-Point Algorithms

Let $F : \mathbb{R}^p \mapsto \mathbb{R}^p$ be the fixed-point mapping associated with a given iterative algorithm. That is, iterates $x_0, x_1, x_2, ...$ are generated from the updating scheme

$$x_{k+1} = F(x_k), \quad k = 0, 1, ..., \tag{1}$$

where x_0 is a given initial point for the algorithm. We will let x^* denote a fixed-point of F ; that is, x^* satisfies $x^* = F(x^*)$. If $F'(x^*)$ denotes Jacobian of F at x^*, the iteration scheme (1) is locally convergent to x^* provided that $\rho\{F'(x^*)\} < 1$, where $\rho\{F'(x^*)\}$ denotes the spectral radius (i.e., the largest eigenvalue) of the Jacobian $F'(x^*)$. By local convergence, we mean the following: there exists a $\delta > 0$ such that whenever $||x_k - x^*|| \leq \delta$ the iteration defined by Eq. (1) converges to x^*.

Assuming that F is sufficiently smooth, we can write a Taylor series expansion of $F(x)$ around the fixed point:

$$F(x) = F(x^*) + F'(x^*)(x - x^*) + o\big(||x - x^*||\big).$$

Because $x_{k+1} = F(x_k)$, we can express x_{k+1} as

$$x_{k+1} = F(x^*) + F'(x^*)(x_k - x^*) + o\big(||x_k - x^*||\big),$$

and hence we obtain

$$\frac{||x_{k+1} - x^*||}{||x_k - x^*||} = \frac{||F'(x^*)(x_k - x^*) + o(||x_k - x^*||)||}{||x_k - x^*||}. \tag{2}$$

To characterize the convergence speed of the iteration (1), we first note that any sequence $y_0, y_1, y_2, \ldots$ that converges to a point y^* is said to converge linearly to y^* with rate $r \in (0, 1)$ provided that

$$\lim_{k \longrightarrow \infty} \frac{||y_{k+1} - y^*||}{||y_k - y^*||} = r.$$

Hence, from (2), we can see that the iterative scheme defined by (1) converges linearly with rate equal to the spectral radius of $F'(x^*)$, provided that $F'(x^*) \neq 0$. In the case of the EM algorithm, the linear convergence rate can be expressed as the ratio of the observed information to the complete information.

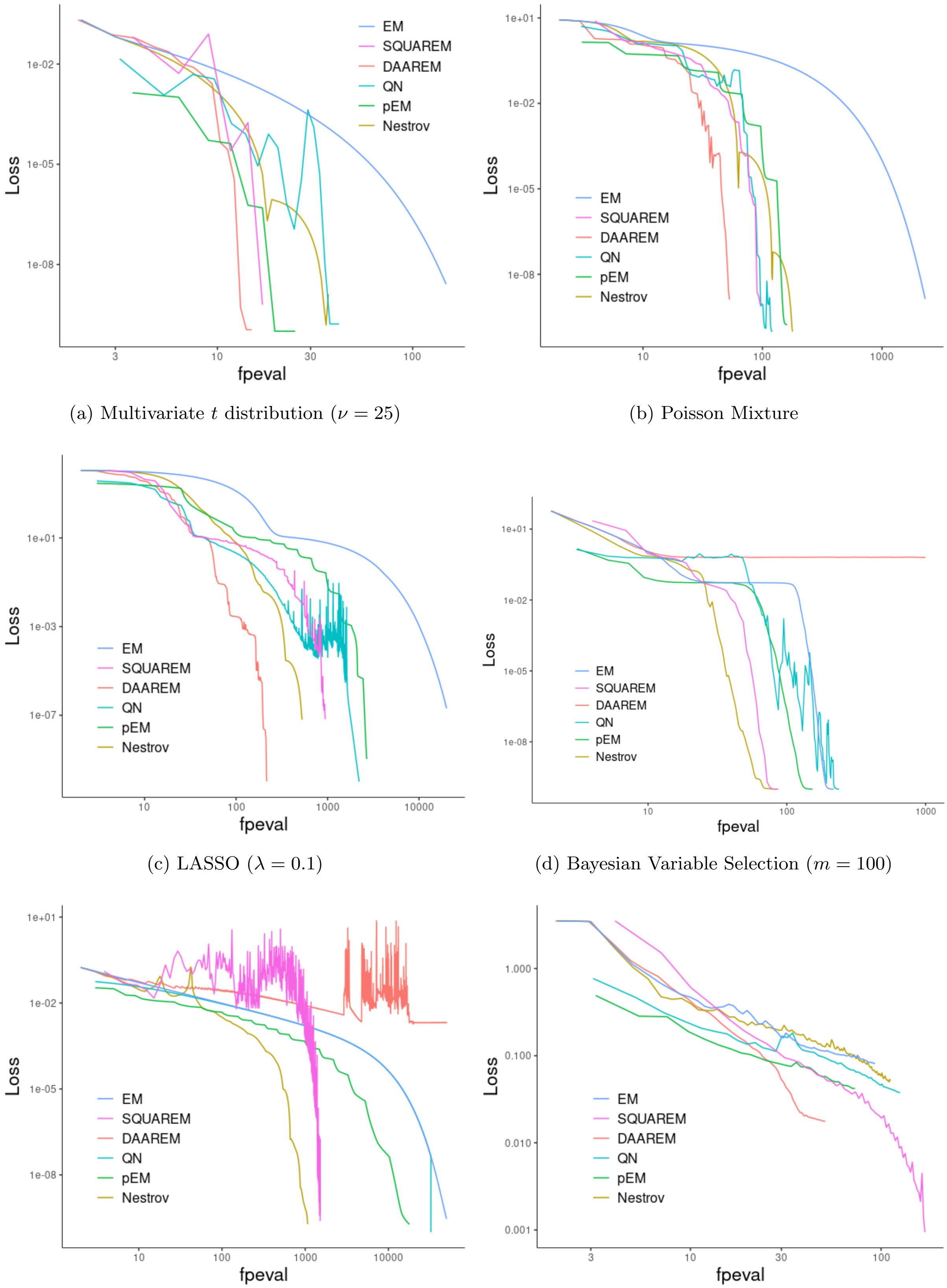
An alternative approach to studying the convergence of iterative processes utilizes Lyapunov functions. [1] shows how this approach can be used to establish the convergence of iterative processes. A mapping $L : \mathbb{R}^p \mapsto \mathbb{R}$ is said to be a Lyapunov function for F at a fixed-point x^* if there is an open neighborhood D around x^* such that

$$\begin{gathered} L(x) > 0, \quad \forall x \in D, x \neq x^*; L(x^*) = 0 \\ L(F(x)) < L(x), \quad \forall\, x, F(x) \in D, x \neq x^* \end{gathered} \tag{3}$$

A fixed-point iteration with fixed point mapping F that is endowed with a Lyapunov function as defined in (3) is guaranteed to be locally convergent. Note that this characterization does not require the existence of a Jacobian of F. In fact, the existence of a Lyapunov function guarantees that the iteration is asymptotically stable, which is a stronger property than local convergence. For the EM algorithms in statistics, a Lyapunov function can be readily constructed from the log-likelihood function, and a Lyapunov function for any MM algorithm can also be readily obtained from the objective function of that problem. The Lyapunov function plays an important role in acceleration algorithms which are generally non-monotone. In the implementation of an acceleration scheme, it is important to enforce some degree of monotonicity after each acceleration step, either by employing some control on the degree of extrapolation or by forsaking extrapolation and relying upon the base algorithm.

2 Convergence Visualization

We give a visualization of convergence for different algorithms in different problems in a typical run, where losses are plotted out against the number of fixed point iteration evaluated until convergence. Please check Figure 1.



(a) Multivariate t distribution ($\nu = 25$)

(b) Poisson Mixture

(c) LASSO ($\lambda = 0.1$)

(d) Bayesian Variable Selection ($m = 100$)

(e) Sinkhorn Scaling (H_{50})

(f) t-SNE

Figure 1: Visualization of convergence. *Loss* is normalized by subtracting the minimum value of the objective function attained across all methods. The figure only shows results from the most difficult setting of each experiment (if there are multiple settings). And **EM** line always indicates the original method that is not accelerated.

3 Sinkhorn Iteration Supplementary Analysis

It is also possible to use a different approach to matrix scaling. Here we treat the intermediate scaled matrix Γ_k as parameters, in which case the Sinkhorn iterations can be write as $\Gamma_{k+1} = D_k \Gamma_k E_k$, with $\text{diag}(D_k) = \boldsymbol{a} \ / \ \Gamma_k \mathbf{1}$ and $\text{diag}(E_k) = \boldsymbol{b} \ / \ \Gamma_k^T \mathbf{1}$. It may be observed that this algorithm is well-known in statistics as the *iterative proportional fitting* (IPF) algorithm, originally proposed by [2]. We accelerate the IPF algorithm where the scaled matrix is treated as the parameter vector to be estimated. We do observe dramatic acceleration based on this base iteration. Table 1 shows the results for order 50 Hessenberg Matrix. It should be noted that this method does not guarantee that the converged Γ obtained from acceleration methods is also a scaling of the original matrix, which means $\Gamma \neq DAE$ could happen for any D and E. There is a discernible deviation from the unique true solution. However, the deviation is fairly small, see the relative difference $\|\Gamma - \Gamma_{true}\|_F \ / \ \|\Gamma_{true}\|_F$ in Table 1 for reference. Therefore, the approach seems to be adequate. It would be an interesting future study to investigate whether acceleration with Γ_k is guaranteed to yield bounded error, and hence might be suitable as a fast approximation algorithm.

Metric	SK	SQUAREM	DAAREM	pEM	Quasi-Newton	Nesterov
fpevals	50000+	461 ± 81.2	$\mathbf{346 \pm 31.5}$	17836 ± 33.6	9606 ± 16782	1006 ± 53.8
elapsed	9.6 ± 0.36	$\mathbf{0.056 \pm 0.015}$	0.114 ± 0.02	2.9 ± 0.19	2.04 ± 3.66	0.14 ± 0.03
rel. diff. $(\times 10^{-6})$	**0**	0.33 ± 0.089	9.53 ± 2.05	0.43 ± 0.071	145.92 ± 266.3	25.88 ± 1.32

Table 1: Scaling $\text{diag}\{\boldsymbol{u}_0\}\boldsymbol{H}_{\mathbf{50}}\text{diag}\{\boldsymbol{v}_0\}$ for 200 independent runs, where $\boldsymbol{u}_0, \boldsymbol{v}_0$ i.i.d drawn from $Unif[0.5, 2]$ to create some randomness. *SK* represents the original Sinkhorn-Knopp algorithm, and the other columns are different accelerated versions of it. Elapsed time are reported in seconds. Rel. diff. refers to the Frobenius norm between the output and the optimal result divided by the Frobenius norm of the optimal result. It is reported in the scale of 10^{-6}. All other settings remain the same as in the main paper.

References

[1] James M Ortega. Stability of difference equations and convergence of iterative processes. *SIAM Journal on Numerical Analysis*, 10(2):268–282, 1973.

[2] W E Deming and F F Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, pages 427–444, 1940.