**Expectation-Maximization (EM)** is a popular tool in statistics and machine learning, first introduced in the 1970s.

- **Applications:** fit models with latent or hidden variables, hidden Markov models, semi-supervised learning, generative models with missing data, etc.
- **Prior works analyzing convergence rate of EM** make very strong assumptions
  - Initial estimate of parameters needs to be close to the optima.
  - Fraction of missing information needs to be small.
  - Other regularity conditions.

**This Work:** We provide a bound on the number of iterations of EM.

- **Provide a lower bound on the decrease in the negative log-likelihood (NLL) on each iteration.**
- **Provide the first convergence rate for non-convex functions in a generalized surrogate optimization framework and, consequently, for EM.**

**Surrogate Optimization**

- Consider the following problem: suppose $A \subseteq \mathbb{R}^d$ is convex, and $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is continuous and bounded below; solve for $\lambda^* \in \arg\min_{\lambda \in A} f(\lambda)$.

- We first generalize the definition of first-order surrogate functions [1].

**Definition**

Let $f$ and $g$ be functions from $\mathbb{R}^d \rightarrow \mathbb{R}$. We say that $g$ is a surrogate of $f$ near $\lambda^k \in A$ if it satisfies:

- **Majorization:** $\forall \lambda \in \arg\min_{\lambda \in A} g(\lambda), f(\lambda) \leq g(\lambda)$. If $g(\lambda) \leq f(\lambda)$ for all $\lambda \in A$, then $g$ is a majorant function;
- **Smoothness:** Denote the approximation error as $h = g - f$. Then, the functions agree at $\lambda^k$ so that $h(\lambda^k) = 0$.

$g_k$ is a majorant surrogate of $f$ at $\lambda^{k-1}$

\[ g_k \triangleq \min_{\lambda \in \Lambda} g(\lambda) \]

Let $g_k$ be a majorant surrogate of $f$ near $\lambda^{k-1}$.

In this setting, Mairal [1] defines the following surrogate optimization framework:

- Initialize parameters $\lambda_0$.
- Compute surrogate $g_k$ of $f$ near $\lambda^{k-1}$.
- Update parameters $\lambda^k \in \arg\min_{\lambda \in A} g_k(\lambda)$.

In contrast to [1], we do not require differentiability of $h_k$ or that $\nabla h_k(\lambda^k) = 0$.

**EM as a Surrogate Optimization Algorithm**

- In EM, we want to find parameters $\lambda \in A$ to maximize the likelihood, $P(X|\lambda)$.
- Introducing hidden or latent variables, we can write the likelihood as $\sum P(X, z|\lambda)$.
- Equivalently, we can minimize the negative log-likelihood (NLL). So, our goal is to find $\lambda^* \in \arg\min_{\lambda \in A} \log \sum P(X, z|\lambda)$.

Let $\lambda^k$ denote the estimate of the parameters after the $k$th iteration and define $Q(\lambda^k) = \sum_{z} P(z, X|\lambda^k) \log P(X|\lambda^k)$.

Using Jensen’s inequality, we get the following well-known upper bound on the NLL

\[ \log P(X|\lambda) \leq -Q(\lambda^k) - \text{ent}(z|X, \lambda^k) \quad \tag{1} \]

The iterations of EM are defined as

\[ \lambda^{k+1} \in \arg\min_{\lambda \in A} -Q(\lambda^k) - \text{ent}(z|X, \lambda^k) \]

Define

\[ f(\lambda) = -\log P(X|\lambda) = -\log \sum_{z} P(X, z|\lambda), \]

\[ g_k(\lambda) = -Q(\lambda^{k-1}) - \text{ent}(z|X, \lambda^{k-1}). \]

We need to verify that $g_k$ as defined above is indeed a surrogate of $f$.

- From equation (1), we see that $g_k$ is a majorant of $f$, and thus, it satisfies the majorization condition.
- It is a well-known fact that $h_k(\lambda^{k-1}) = 0$, and thus, it satisfies the smoothness condition.

In addition, to derive our convergence results, we will assume that for all iterations, $g_k$ is $\rho$-strongly-convex. This is satisfied in many scenarios, like mixtures of exponential families, or when using a strongly-convex regularizer with a convex complete-data NLL.

**Convergence rate**

- Informally, if the iterates stay within a convex set and the surrogates are $\rho$-strongly convex, then the further away successive iterates are, the greater the decrease in the objective.

**Theorem**

Let $g_k$ be a $\rho$-strongly-convex surrogate of $f$ near $\lambda^{k-1}$, and $\lambda^k \in \arg\min_{\lambda \in A} g_k(\lambda)$.

Then,

\[ f(\lambda^k) - f(\lambda^{k-1}) \leq \frac{\rho}{2} \| \lambda^k - \lambda^{k-1} \|^2 \quad \tag{2} \]

**Proof**

Using that $\lambda^k$ minimizes $g_k$ and that $g_k$ is $\rho$-strongly-convex, it follows that for all $\lambda \in A$,

\[ g_k(\lambda) = P(\lambda) - \frac{\rho}{2} \| \lambda - \lambda^k \|^2 \leq g_k(\lambda^k), \]

Now using that $g_k$ is a majorant, we get

\[ f(\lambda^k) = f(\lambda^k) + \frac{\rho}{2} \| \lambda^k - \lambda^{k-1} \|^2 \leq g_k(\lambda^k) + \frac{\rho}{2} \| \lambda^k - \lambda^{k-1} \|^2 \]

\[ f(\lambda^k) \leq \frac{\rho}{2} \| \lambda^k - \lambda^{k-1} \|^2 \leq f(\lambda^{k-1}) - f(\lambda^k), \]

which can be re-arranged to get the result.

- We use this bound to derive an $O(\frac{1}{t})$ convergence rate in terms of the squared difference between successive iterates.

**Theorem**

Let $g_k$ be a $\rho$-strongly-convex surrogate of $f$ near $\lambda^{k-1}$, and $\lambda^k \in \arg\min_{\lambda \in A} g_k(\lambda)$.

Then,

\[ \min_{\lambda \in A} \| \lambda - \lambda^{k-1} \|^2 \leq \frac{2(f(\lambda^k) - f(\lambda^{k-1}))}{\rho t} \quad \tag{4} \]

**Proof**

Summing up (3) for all $k$ and telescoping the sum we get

\[ \sum_{k=1}^t \frac{1}{2} \| \lambda^k - \lambda^{k-1} \|^2 \leq \sum_{k=1}^t \| f(\lambda^k) - f(\lambda^{k-1}) \|^2 \leq \frac{f(\lambda^k) - f(\lambda^{k-1})}{\rho t} \]

Taking the min over all iterations, we get

\[ \min_{\lambda \in A} \| \lambda - \lambda^{k-1} \|^2 \leq \frac{2(f(\lambda^k) - f(\lambda^{k-1}))}{\rho t} 

**Discussion**

- Our analysis is quite general and relies on mild assumptions.
- If we make a slightly stronger assumption that the approximation error $h_k$ is differentiable, $\nabla h_k$ is $\rho$-Lipschitz continuous, and the gradients agree, i.e., $\nabla g_k(\lambda^{k-1}) = 0$, then we can derive a similar convergence rate in terms of the norm of the gradient of $f$.
- Using the above, the standard gradient descent progress bound and that $\lambda^k$ is a global minimizer of $g_k$, we can follow the above proofs to derive

\[ \min_{\lambda \in A} \| \nabla f(\lambda^k) \|^2 \leq \frac{2f(\lambda^k) - f(\lambda^{k-1})}{\rho t} \]

**Future work**

- It would be interesting to see if some assumptions could be relaxed, like strong-convexity of the surrogates.
- It would also be interesting to derive stronger convergence results using the same set of assumptions for “nice” scenarios, like mixtures of exponential families.
- Viewing EM in such an optimization framework allows future work to use numerical optimization techniques to develop improved variants of EM, like a variance reduced version of EM.

**References**