Convergence rate of expectation-maximization

Raunak Kumar (UBC), Mark Schmidt (UBC)

Expectation-Maximization

- 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 $\Lambda \subset \mathbb{R}^d$ is convex, and $f : \mathbb{R}^d \to \mathbb{R}$ is continuous and bounded below; solve for

$$\lambda^* \in \operatorname*{argmin}_{\lambda \in \Lambda} f(\lambda).$$

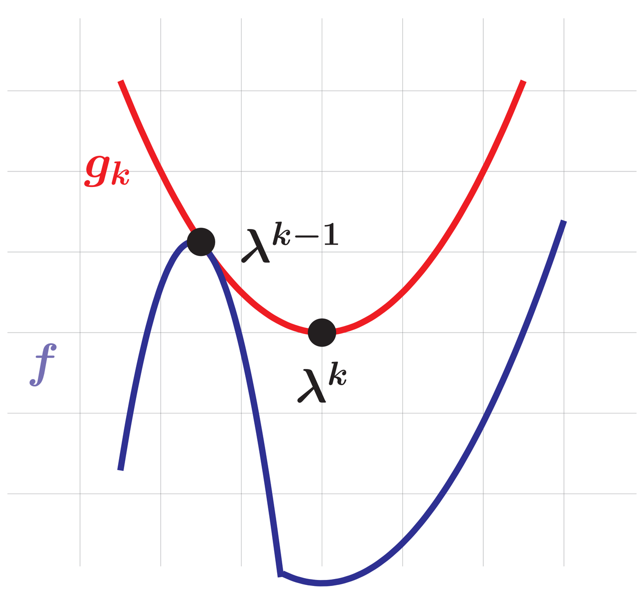
- We first generalize the definition of first-order surrogate functions [1].

Definition

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

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

g_k is a majorant surrogate of f at λ^{k-1}



- In this setting, Mairal [1] defines the following surrogate optimization framework:
  - → Initialize parameters λ^0.
  - → Compute surrogate g_k of f near λ^{k-1}.
  - → Update parameters $\lambda^k \in \operatorname{argmin}_{\lambda\in\Lambda} g(\lambda)$.
- 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 $\lambda \in \Lambda$ to maximize the likelihood, $P(X|\lambda)$.
- Introducing hidden or latent variables, we can write the likelihood as $\sum_z P(X, z|\lambda)$.
- Equivalently, we can minimize the negative log-likelihood (NLL). So, our goal is to find

$$\lambda^* \in \operatorname*{argmin}_{\lambda\in\Lambda} -\log \sum_z P(X, z|\lambda).$$

- Let λ^k denote the estimate of the parameters after the k^{th} iteration and define

$$Q(\lambda|\lambda^k) = \sum_z P(z|X, \lambda^k) \log P(X, z|\lambda).$$

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

$$-\log P(X|\lambda) \leq -Q(\lambda|\lambda^k) - \text{entropy}(z|X, \lambda^k). \tag{1}$$

- The iterations of EM are defined as

$$\begin{aligned}\lambda^{k+1} &\in \operatorname*{argmin}_{\lambda\in\Lambda} -Q(\lambda|\lambda^k) - \text{entropy}(z|X, \lambda^k)\\ &\equiv \lambda^{k+1} \in \operatorname*{argmin}_{\lambda\in\Lambda} -Q(\lambda|\lambda^k).\end{aligned}$$

- Define

$$f(\lambda) = -\log P(X|\lambda) = -\log \sum_z P(X, z|\lambda),$$

$$g_k(\lambda) = -Q(\lambda|\lambda^{k-1}) - \text{entropy}(z|X, \lambda^{k-1}).$$

- We need to verify that g_k as defined above is indeed a surrogate of f.
  - → From equation (1), we can 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 ρ-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

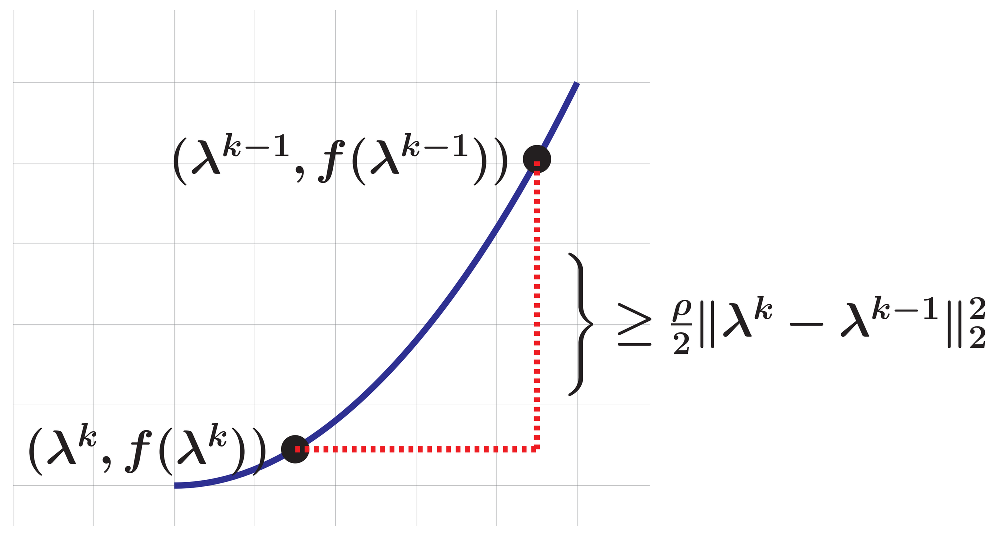
- Informally, if the iterates stay within a convex set and the surrogates are ρ-strongly convex, then the further away successive iterates are, the greater the decrease in the objective.

Theorem

Let g_k be a ρ-strongly-convex surrogate of f near λ^{k-1}, and $\lambda^k \in \operatorname{argmin}_{\lambda\in\Lambda} g_k(\lambda)$. Then,

$$f(\lambda^k) \leq f(\lambda^{k-1}) - \frac{\rho}{2}\|\lambda^k - \lambda^{k-1}\|_2^2. \tag{2}$$

Lower bound on decrease in NLL



Proof.
Using that λ^k minimizes g_k and that g_k is ρ-strongly-convex, it follows that for all $\lambda \in \Lambda$,

$$g_k(\lambda^k) + \frac{\rho}{2}\|\lambda - \lambda^k\|_2^2 \leq g_k(\lambda).$$

Now using that g_k is a majorant, we get

$$\begin{aligned}f(\lambda^k) + \frac{\rho}{2}\|\lambda^k - \lambda\|_2^2 &\leq g_k(\lambda^k) + \frac{\rho}{2}\|\lambda^k - \lambda\|_2^2\\ &\leq g_k(\lambda)\\ &= f(\lambda) + h_k(\lambda).\end{aligned}$$

Setting $\lambda = \lambda^{k-1}$ and using that $h_k(\lambda^{k-1}) = 0$ from the definition of surrogate functions gives

$$\begin{aligned}f(\lambda^k) + \frac{\rho}{2}\|\lambda^k - \lambda^{k-1}\|_2^2 &\leq f(\lambda^{k-1}) + h_k(\lambda^{k-1})\\ \frac{\rho}{2}\|\lambda^k - \lambda^{k-1}\|_2^2 &\leq f(\lambda^{k-1}) - f(\lambda^k),\end{aligned} \tag{3}$$

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 ρ-strongly-convex surrogate of f near λ^{k-1}, and $\lambda^k \in \operatorname{argmin}_{\lambda\in\Lambda} g_k(\lambda)$. Then,

$$\min_{k\in\{1,2,\ldots,t\}} \|\lambda^k - \lambda^{k-1}\|_2^2 \leq \frac{2(f(\lambda^0) - f(\lambda^*))}{\rho t}. \tag{4}$$

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

$$\begin{aligned}\sum_{k=1}^t \frac{\rho}{2}\|\lambda^k - \lambda^{k-1}\|_2^2 &\leq \sum_{k=1}^t f(\lambda^{k-1}) - f(\lambda^k)\\ &= f(\lambda^0) - f(\lambda^t)\\ &\leq f(\lambda^0) - f(\lambda^*).\end{aligned}$$

Taking the min over all iterations, we get

$$\min_{k\in 1,2,\ldots,t} \|\lambda^k - \lambda^{k-1}\|_2^2 \cdot \frac{\rho t}{2} \leq f(\lambda^0) - f(\lambda^*)$$

$$\min_{k\in\{1,2,\ldots,t\}} \|\lambda^k - \lambda^{k-1}\|_2^2 \leq \frac{2(f(\lambda^0) - f(\lambda^*))}{\rho t}.$$

□

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, ∇h_k is L-Lipschitz continuous, and the gradients agree, ie. $\nabla h(\lambda^{k-1}) = 0$, 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 λ^k is a global minimizer of g_k, we can follow the above proofs to derive

$$\min_{k\in\{1,2,\ldots,t\}} \|\nabla f(\lambda^{k-1})\|_2^2 \leq \frac{2L(f(\lambda^0) - f(\lambda^*))}{t}. \tag{5}$$

- 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

[1] Mairal, J., 2013. Optimization with first-order surrogate functions. In Proceedings of the 30th International Conference on Machine Learning (ICML-13) (pp. 783-791).