PALM: Probabilistic Area Loss Minimization for Protein Sequence Alignment (Supplementary Materials)

Fan Ding[*1] **Nan Jiang**[*1] **Jianzhu Ma**[2] **Jian Peng**[3] **Jinbo Xu**[4] **Yexiang Xue**[1]

[1]Department of Computer Science, Purdue University, West Lafayette, Indiana, USA
[2]Institute for Artificial Intelligence, Peking University, Beijing, China
[3]Department of Computer Science, University of Illinois at Urbana-Champaign, Illinois, USA
[4]Toyota Technological Institute at Chicago, Illinois, USA

1 PROOF OF THEOREM 1

Theorem 1 states the function value of the output of PALM, in expectation converges to the true optimum within a small constant distance at a linear speed w.r.t. the number of iterations T. To prove Theorem 1, we need the following lemma.

Lemma 1. *If the total variation* $\max_\theta Var_{P_\theta}(\phi(a)) \leq L$*, then* $l(\theta)$ *is* L*-smooth w.r.t.* θ*.*

1.1 PROOF OF LEMMA 1

Proof. L-smoothness requires that

$$\|\nabla \mathcal{L}_{LB}(\theta_1) - \nabla \mathcal{L}_{LB}(\theta_2)\|_2 \leq L\|\theta_1 - \theta_2\|_2,$$

where $\forall \theta_1, \theta_2 \in \ dom\ f$ and L is a constant. Based on the mean value theorem, there exists a point $\tilde{\theta} \in (\theta_1, \theta_2)$ such that

$$\nabla \mathcal{L}_{LB}(\theta_1) - \nabla \mathcal{L}_{LB}(\theta_2) = \nabla(\nabla \mathcal{L}_{LB}(\tilde{\theta}))(\theta_1 - \theta_2).$$

Taking the L_2 norm for both sides, we have

$$\|\nabla \mathcal{L}_{LB}(\theta_1) - \nabla \mathcal{L}_{LB}(\theta_2)\|_2 = \|\nabla(\nabla \mathcal{L}_{LB}(\tilde{\theta}))(\theta_1 - \theta_2)\|_2 \leq \|\nabla(\nabla \mathcal{L}_{LB}(\tilde{\theta}))\|_2 \ \|\theta_1 - \theta_2\|_2$$

Then, the problem is to bound the matrix 2-norm $\|\nabla(\nabla \mathcal{L}_{LB}(\tilde{\theta}))\|_2$. Since we know the explicit form of $\mathcal{L}_{LB}(\theta)$, we know

$$\begin{aligned}\nabla \mathcal{L}_{LB}(\theta) &= \nabla \log Z_\phi - \phi(a),\\ \nabla(\nabla \mathcal{L}_{LB}(\theta)) &= \sum_a [\phi(a) - \nabla \log Z_\phi][\phi(a) - \nabla \log Z_\phi]^T P_\theta(a),\end{aligned}$$

where $\nabla(\nabla \mathcal{L}_{LB}(\theta))$ is the co-variance matrix. Denote $\text{Cov}_\theta[\phi(a)] = \nabla(\nabla \mathcal{L}_{LB}(\theta))$, which is both symmetric and positive semi-definite. We have

$$\|\nabla(\nabla \mathcal{L}_{LB}(\tilde{\theta}))\|_2 = \|\text{Cov}_\theta[\phi(a)]\|_2 = \lambda_{\max},$$

where $\lambda_{\max}$ is the maximum eigenvalue of the matrix $\text{Cov}_\theta[\phi(a)]$. Then, because of the positive semi-definiteness of the co-variance matrix, all the eigenvalues are non-negative, and we can bound λ_{max} as

$$\lambda_{max} \leq \sum_i \lambda_i = Tr(\text{Cov}_\theta[\phi(a)]),$$

[*]These authors contribute equally.

where $Tr(\text{Cov}_\theta[\phi(a)])$ is the trace of matrix $\text{Cov}_\theta[\phi(a)]$. $Tr(\text{Cov}_\theta[\phi(a)])$ can be further derived as:

$$Tr(\text{Cov}_\theta[\phi(a)]) = \mathbb{E}_{P_\theta}[\|\phi(a)\|_2^2] - \|\mathbb{E}_{P_\theta}[\phi(a)]\|_2^2,$$

which is equal to the total variation $Var_{P_\theta}(\phi(a))$, we have

$$\|\nabla(\nabla\mathcal{L}_{LB}(\tilde{\theta}))\|_2 \leq Var_{P_\theta}(\phi(X)) \leq L.$$

Therefore, we have

$$\|\nabla l(\theta_1) - \nabla l(\theta_2)\|_2 \leq L\,\|\theta_1 - \theta_2\|_2.$$

This completes the proof. □

1.2 PROOF OF THEOREM 1

Proof. By L-smooth of $\mathcal{L}_{LB}$, we have for the t-th iteration,

$$\begin{aligned}\mathcal{L}_{LB}(\theta_{t+1}) &\leq \mathcal{L}_{LB}(\theta_t) + \langle\nabla\mathcal{L}_{LB}(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|_2^2,\\ &= \mathcal{L}_{LB}(\theta_t) - \eta\langle\nabla\mathcal{L}_{LB}(\theta_t), g_t\rangle + \frac{L\eta^2}{2}\|g_t\|^2.\end{aligned}$$

Because of $\mathbb{E}[g_t]^2 = \mathbb{E}[\|g_t\|_2^2] - Var(g_t)$, by taking expectation on both sides w.r.t g_t we get

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{LB}(\theta_{t+1})] = \mathcal{L}_{LB}(\theta_t) - \eta\mathbb{E}[g_t]^2 + \frac{L\eta^2}{2}\mathbb{E}[\|g_t\|_2^2] &= \mathcal{L}_{LB}(\theta_t) - \eta(\mathbb{E}[\|g_t\|_2^2] - Var(g_t)) + \frac{L\eta^2}{2}\mathbb{E}[\|g_t\|_2^2],\\ &\leq \mathcal{L}_{LB}(\theta_t) - \eta(1 - \frac{L\eta}{2})\mathbb{E}[\|g_t\|_2^2] + \frac{\eta\sigma^2}{M},\\ &\leq \mathcal{L}_{LB}(\theta_t) - \frac{\eta}{2}\mathbb{E}[\|g_t\|_2^2] + \frac{\eta\sigma^2}{M}.\end{aligned}$$

where the last inequality follows as $L\eta \leq 2$. Because $\mathcal{L}_{LB}$ is convex, we get

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{LB}(\theta_{t+1})] &\leq \mathcal{L}_{LB}(\theta^*) + \langle\nabla\mathcal{L}_{LB}(\theta_t), \theta_t - \theta^*\rangle - \frac{\eta}{2}\mathbb{E}[\|g_t\|_2^2] + \eta\sigma^2,\\ &= \mathcal{L}_{LB}(\theta^*) + \langle\mathbb{E}[g_t], \theta_t - \theta^*\rangle - \frac{\eta}{2}\mathbb{E}[\|g_t\|_2^2] + \frac{\eta\sigma^2}{M},\\ &= \mathcal{L}_{LB}(\theta^*) + \mathbb{E}[\langle g_t, \theta_t - \theta^*\rangle - \frac{\eta}{2}\|g_t\|_2^2] + \frac{\eta\sigma^2}{M}.\end{aligned}$$

we now repeat the calculations by completing the square for the middle two terms to get

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{LB}(\theta_{t+1})] &\leq \mathcal{L}_{LB}(\theta^*) + \mathbb{E}[\frac{1}{2\eta}(\|\theta_t - \theta^*\|_2^2 - \|\theta_t - \theta^* - \eta g_t\|_2^2)] + \frac{\eta\sigma^2}{M},\\ &= \mathcal{L}_{LB}(\theta^*) + \mathbb{E}[\frac{1}{2\eta}(\|\theta_t - \theta^*\|_2^2 - \|\theta_{t+1} - \theta^*\|_2^2)] + \frac{\eta\sigma^2}{M}.\end{aligned}$$

Summing the above equations for $t = 0, \ldots, T-1$, we get

$$\sum_{t=0}^{T-1}\mathbb{E}[\mathcal{L}_{LB}(\theta_{t+1}) - \mathcal{L}_{LB}(\theta^*)] \leq \frac{1}{2\eta}(\|\theta_0 - \theta^*\|_2^2 - \mathbb{E}[\|\theta_T - \theta^*\|_2^2]) + T\frac{\eta\sigma^2}{M} \leq \frac{\|\theta_0 - \theta^*\|_2^2}{2\eta} + T\frac{\eta\sigma^2}{M}.$$

Finally, by Jensen's inequality, $T\mathcal{L}_{LB}(\overline{\theta_T}) \leq \sum_{t=1}^T \mathcal{L}_{LB}(\theta_t)$, thus,

$$\sum_{t=0}^{T-1}\mathbb{E}[\mathcal{L}_{LB}(\theta_{t+1}) - \mathcal{L}_{LB}(\theta^*)] = \mathbb{E}[\sum_{t=1}^{T}\mathcal{L}_{LB}(\theta_t)] - T\mathcal{L}_{LB}(\theta^*) \geq T\mathbb{E}[\mathcal{L}_{LB}(\overline{\theta_T})] - T\mathcal{L}_{LB}(\theta^*).$$

Combining the above equations we get

$$\mathbb{E}[\mathcal{L}_{LB}(\overline{\theta_T})] \leq \mathcal{L}_{LB}(\theta^*) + \frac{\|\theta_0 - \theta^*\|_2^2}{2\eta T} + \frac{\eta\sigma^2}{M}.$$

This completes the proof. □