Appendix

A Strong convexity

As we discussed, the posterior from Bayes's rule could be viewed as the optimal of an optimization problem in Eq (1). We will show that the objective function is strongly convex w.r.t KL-divergence.

Proof for Lemma 1. The lemma directly results from the generalized Pythagaras theorem for Bregman divergence. Particularly, for KL-divergence, we have

$$KL(q_1||q) \quad = \quad KL(q_1||q_2) + KL(q_2||q) - \langle q_1 - q_2, \nabla\phi(q) - \nabla\phi(q_2)\rangle_2$$

where $\phi(q)$ is the entropy of q.

Notice that $L(q) = KL(q||q^*) - \log Z$, where $q^* = \frac{p(\theta)\Pi_i^N p(x_i|\theta)}{Z}$, $\quad Z = \int p(\theta)\Pi_i^N p(x_i|\theta)$, we have

$$\begin{aligned} & KL(q_1||q^*) - KL(q_2||q^*) - \langle q_1 - q_2, \nabla\phi(q_2) - \nabla\phi(q^*)\rangle_2 = KL(q_1||q_2) \\ \Rightarrow\quad & KL(q_1||q^*) - KL(q_2||q^*) - \langle q_1 - q_2, \log q_2 - \log q^*\rangle_2 = KL(q_1||q_2) \\ \Rightarrow\quad & KL(q_1||q^*) - KL(q_2||q^*) - \langle q_1 - q_2, \log q_2 - \log\left(p(\theta)\Pi_i^N p(x_i|\theta)\right)\rangle_2 + \underbrace{\langle q_1 - q_2, \log Z\rangle_2}_{0} = KL(q_1||q_2) \\ \Rightarrow\quad & L(q_1) - L(q_2) - \langle q_1 - q_2, \nabla L(q_2)\rangle_2 = KL(q_1||q_2) \end{aligned}$$

■

B Finite Convergence of Stochastic Mirror Descent with Inexact Prox-Mapping in Density Space

Since the prox-mapping of stochastic mirror descent is intractable when directly being applied to the optimization problem (1), we propose the ϵ*-inexact prox-mapping* within the stochastic mirror descent framework in Section 3. Instead of solving the prox-mapping exactly, we approximate the solution with ϵ error. In this section, we will show as long as the approximation error is tolerate, the stochastic mirror descent algorithm still converges.

Theorem 2 *Denote $q^* = \mathrm{argmin}_{q\in\mathcal{P}} L(q)$, the stochastic mirror descent with inexact prox-mapping after T steps gives*

(a) the recurrence: $\forall t \leqslant T$, $\mathbb{E}[KL(q^*||\tilde{q}_{t+1})] \leqslant \epsilon_t + (1-\gamma_t)\mathbb{E}[KL(q^*||\tilde{q}_t)] + \frac{\gamma_t^2\mathbb{E}\|g_t\|_\infty^2}{2}$

(b) the sub-optimality: $\mathbb{E}[KL(\bar{q}_T||q^*)] \leqslant \mathbb{E}[L(\bar{q}_T) - L(q^*)] \leqslant \frac{\mathcal{M}^2\cdot\frac{1}{2}\sum_{t=1}^T\gamma_t^2 + \sum_{t=1}^T\epsilon_t + D_1}{\sum_{t=1}^T\gamma_t}$ where $\bar{q}_T = \sum_{t=1}^T\gamma_t\tilde{q}_t/\sum_{t=1}^T\gamma_t$ and $D_1 = KL(q^*||\tilde{q}_1)$ and $\mathcal{M}^2 := \max_{1\leqslant t\leqslant T}\mathbb{E}\|g_t\|_\infty^2$.

Remark. Based on (Nemirovski et al., 2009), one can immediately see that, to guarantee the usual rate of convergence, the error ϵ_t can be of order $O(\gamma_t^2)$. The first recurrence implies an overall $O(1/T)$ rate of convergence for the KL-divergence when the stepsize γ_t is as small as $O(1/t)$ and error ϵ_t is as small as $O(1/t^2)$. The second result implies an overall $O(1/\sqrt{T})$ rate of convergence for objective function when larger stepsize $\gamma_t = O(1/\sqrt{T})$ and larger error $\epsilon_t = O(1/t)$ are adopted.

Proof for Theorem 2. (a) By first-order optimality condition, $\tilde{q}_{t+1} \in P_{\tilde{q}_t}^{\epsilon_t}(\gamma_t g_t)$ is equivalent as

$$\langle\gamma_t g_t + \log(\tilde{q}_{t+1}) - \log(\tilde{q}_t), \tilde{q}_{t+1} - q\rangle_{L_2} \leqslant \epsilon_t, \forall q \in \mathcal{P},$$

which implies that

$$\langle\gamma_t g_t, \tilde{q}_{t+1} - q\rangle_2 \leqslant \langle\log(\tilde{q}_t) - \log(\tilde{q}_{t+1}), \tilde{q}_{t+1} - q\rangle_2 + \epsilon_t = KL(q||\tilde{q}_t) - KL(q||\tilde{q}_{t+1}) - KL(\tilde{q}_{t+1}||\tilde{q}_t) + \epsilon_t$$

Hence,

$$\langle\gamma_t g_t, \tilde{q}_t - q\rangle_2 \leqslant KL(q||\tilde{q}_t) - KL(q||\tilde{q}_{t+1}) - KL(\tilde{q}_{t+1}||\tilde{q}_t) + \langle\gamma_t g_t, \tilde{q}_t - \tilde{q}_{t+1}\rangle_2 + \epsilon_t. \tag{8}$$

By Young's inequality, we have

$$\langle\gamma_t g_t, \tilde{q}_t - \tilde{q}_{t+1}\rangle_2 \leqslant \frac{1}{2}\|\tilde{q}_t - \tilde{q}_{t+1}\|_1^2 + \frac{\gamma_t^2}{2}\|g_t\|_\infty^2. \tag{9}$$

Also, from Pinsker's inequality, we have

$$KL(\tilde{q}_{t+1}||\tilde{q}_t) \geqslant \frac{1}{2}\|\tilde{q}_t - \tilde{q}_{t+1}\|_1^2. \tag{10}$$

Therefore, combining (8), (9), and (10), we have $\forall q \in \mathcal{P}$

$$\langle \gamma_t g_t, \tilde{q}_t - q\rangle_2 \leqslant \epsilon_t + KL(q||\tilde{q}_t) - KL(q||\tilde{q}_{t+1}) + \frac{\gamma_t^2}{2}\|g_t\|_\infty^2$$

Plugging q^* and taking expectation on both sides, the LHS becomes

$$\mathbb{E}_x\left[\langle \tilde{q}_t - q^*, \gamma_t g_t\rangle\right] \quad = \quad \mathbb{E}_x\left[\langle \tilde{q}_t - q^*, \gamma_t \mathbb{E}[g_t]\rangle \middle| x_{[t-1]}\right] = \mathbb{E}_x\left[\langle \tilde{q}_t - q^*, \gamma_t \nabla L(\tilde{q}_t)\rangle\right],$$

Therefore, we have

$$\mathbb{E}_x\left[\langle \tilde{q}_t - q^*, \gamma_t \nabla L(\tilde{q}_t)\rangle\right] \leqslant \epsilon_t + \mathbb{E}_x\left[KL(q^*||\tilde{q}_t)\right] - \mathbb{E}_x\left[KL(q^*||\tilde{q}_{t+1})\right] + \frac{\gamma_t^2}{2}\mathbb{E}_x\|g_t\|_\infty^2 \tag{11}$$

Because the objective function is 1-strongly convex w.r.t. KL-divergence,

$$\langle q' - q, \nabla L(q') - \nabla L(q)\rangle = KL(q'||q) + KL(q||q'),$$

and the optimality condition, we have

$$\langle \tilde{q}_t - q^*, \nabla L(\tilde{q}_t)\rangle \geqslant KL(q^*||\tilde{q}_t)$$

we obtain the recursion with inexact prox-mapping,

$$\mathbb{E}_x[KL(q^*||\tilde{q}_{t+1})] \leqslant \epsilon_t + (1-\gamma_t)\mathbb{E}_x[KL(q^*||\tilde{q}_t)] + \frac{\gamma_t^2}{2}\mathcal{M}^2$$

(b) Summing over $t = 1, \ldots, T$ of equation (11), we get

$$\sum_{t=1}^T \mathbb{E}_x[\langle \tilde{q}_t - q^*, \gamma_t \nabla L(\tilde{q}_t)\rangle] \leqslant \sum_{t=1}^T \epsilon_t + KL(q^*||\tilde{q}_1) + \sum_{t=1}^T \frac{\gamma_t^2}{2}\mathcal{M}^2$$

By convexity and optimality condition, this leads to

$$\left(\sum_{t=1}^T \gamma_t\right)\mathbb{E}_x[L(\bar{q}_T) - L(q^*)] \leqslant \mathbb{E}_x\left[\sum_{t=1}^T \gamma_t(L(\tilde{q}_t) - L(q^*))\right] \leqslant \sum_{t=1}^T \epsilon_t + KL(q^*||\tilde{q}_1) + \sum_{t=1}^T \frac{\gamma_t^2}{2}\mathcal{M}^2$$

Furthermore, combined with the 1-strongly-convexity, it immediately follows that

$$\mathbb{E}_x[KL(\bar{q}_T||q^*)] \quad \leqslant \quad \mathbb{E}_x[L(\bar{q}_T) - L(q^*)] \leqslant \frac{\frac{1}{2}\sum_{t=1}^T \gamma_t^2\mathcal{M}^2 + \sum_{t=1}^T \epsilon_t + D_1}{\sum_{t=1}^T \gamma_t}.$$

■

C Convergence Analysis for Integral Approximation

In this section, we provide the details of the convergence analysis of the proposed algorithm in terms of integral approximation w.r.t. the true posterior using a good initialization.

Assume that the prior $p(\theta)$ has support Ω cover true posterior distribution $q^*(\theta)$, then, we could represent

$$q^*(\theta) \in \mathcal{F} = \left\{q(\theta) = \alpha(\theta)p(\theta), \int \alpha(\theta)p(\theta)d\theta = 1, 0 \leqslant \alpha(\theta) \leqslant C\right\}.$$

Therefore, one can show

Lemma 7 *$\forall q \in \mathcal{F}$, let $\{\theta_i\}_{i=1}^m$ is* i.i.d. *sampled from $p(\theta)$, we could construct $\hat{q}(\theta) = \sum_{i=1}^m \frac{\alpha(\theta_i)\delta(\theta_i)}{\sum_i^m \alpha(\theta_i)}$, such that $\forall f(\theta) : \mathbb{R}^d \to \mathbb{R}$ bounded and integrable,*

$$\mathbb{E}\left[\left|\int \hat{q}(\theta)f(\theta)d\theta - \int q(\theta)f(\theta)d\theta\right|\right] \leqslant \frac{2\sqrt{C}\|f\|_\infty}{\sqrt{m}}.$$

Proof

Given $q(\theta)$, we sample *i.i.d.*$\{\theta_i\}_{i=1}^m$ from $p(\theta)$, and construct a function

$$\hat{q}(\theta) = \frac{1}{m}\sum_{i=1}^m \alpha(\theta_i)\delta(\theta_i, \theta).$$

It is obviously that

$$\mathbb{E}_\theta[\hat{q}(\theta)] = \mathbb{E}_\theta\left[\frac{1}{m}\sum_{i=1}^m \alpha(\theta_i)\delta(\theta_i, \theta)\right] = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_\theta\left[\alpha(\theta_i)\delta(\theta_i, \theta)\right] = q(\theta)$$

and

$$\mathbb{E}_\theta\left[\int \hat{q}(\theta)f(\theta)d\theta\right] = \mathbb{E}_\theta\left[\frac{1}{m}\sum_{i=1}^m \alpha(\theta_i)f(\theta_i)\right] = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_\theta\left[\alpha(\theta_i)f(\theta_i)\right] = \int q(\theta)f(\theta)d\theta$$

Then,

$$\begin{aligned}
\mathbb{E}_\theta\left[\left|\int \hat{q}(\theta)f(\theta)d\theta - \int q(\theta)f(\theta)d\theta\right|^2\right] &= \mathbb{E}_\theta\left[\left|\int \hat{q}(\theta)f(\theta)d\theta - \mathbb{E}_\theta\left[\int \hat{q}(\theta)f(\theta)d\theta\right]\right|^2\right] \\
= \frac{1}{m}\left(\mathbb{E}_\theta\|\alpha(\theta_i)f(\theta_i)\|_2^2 - \|\mathbb{E}_\theta[\alpha(\theta_i)f(\theta_i)]\|_2^2\right) &\leqslant \frac{1}{m}\mathbb{E}_\theta\|\alpha(\theta_i)f(\theta_i)\|_2^2 = \frac{1}{m}\int \alpha(\theta)^2 f(\theta)^2\pi(\theta)d\theta \\
= \frac{1}{m}\int \alpha(\theta)f(\theta)^2 q(\theta)d\theta &\leqslant \frac{C}{m}\|f(\theta)\|_\infty^2\int \alpha(\theta)q(\theta)d\theta \leqslant \frac{C}{m}\|f(\theta)\|_\infty^2\|\alpha(\theta)\|_\infty
\end{aligned}$$

By Jensen's inequality, we have

$$\mathbb{E}_\theta\left[\left|\int \hat{q}(\theta)f(\theta)d\theta - \int q(\theta)f(\theta)d\theta\right|\right] \leqslant \sqrt{\mathbb{E}_\theta\left[\left|\int \hat{q}(\theta)f(\theta)d\theta - \int q(\theta)f(\theta)d\theta\right|^2\right]} \leqslant \frac{\sqrt{C}\|f(\theta)\|_\infty}{\sqrt{m}}$$

Apply the above conclusion to $f(\theta) = 1$, we have

$$\mathbb{E}\left[\left|\frac{1}{m}\sum_i^m \alpha_i - 1\right|\right] \leqslant \frac{\sqrt{C}}{\sqrt{m}}$$

Let $\tilde{q}(\theta) = \frac{\sum_i^m \alpha(\theta_i)\delta(\theta_i,\cdot)}{\sum_i^m \alpha(\theta_i)}$, then $\sum_i^m \frac{\alpha_i}{\sum_i^m \alpha_i} = 1$, and

$$\begin{aligned}
&\mathbb{E}_\theta\left[\left|\int \tilde{q}(\theta)f(\theta)d\theta - \int \hat{q}(\theta)f(\theta)d\theta\right|\right] = \mathbb{E}_\theta\left[\left|\frac{1}{\sum_i^m \alpha(\theta_i)}\sum_i^m \alpha(\theta_i)f(\theta_i) - \frac{1}{m}\sum_i^m \alpha(\theta_i)f(\theta_i)\right|\right] \\
= \quad &\mathbb{E}_\theta\left[\left|1 - \frac{\sum_i^m \alpha(\theta_i)}{m}\right|\left|\frac{1}{\sum_i^m \alpha(\theta_i)}\sum_i^m \alpha(\theta_i)f(\theta_i)\right|\right] \\
= \quad &\mathbb{E}_\theta\left[\left|1 - \frac{\sum_i^m \alpha(\theta_i)}{m}\right|\frac{1}{\sum_i^m \alpha(\theta_i)}\sum_i^m \alpha(\theta_i)|f(\theta_i)|\right] \leqslant \mathbb{E}\left[\left|1 - \frac{\sum_i^m \alpha_i}{m}\|f(\theta)\|_\infty\right|\right] \leqslant \frac{\sqrt{C}\|f(\theta)\|_\infty}{\sqrt{m}}
\end{aligned}$$

Then, we have achieve our conclusion that

$$\begin{aligned}
&\mathbb{E}_\theta\left[\left|\int \tilde{q}(\theta)f(\theta)d\theta - \int q(\theta)f(\theta)d\theta\right|\right] \\
\leqslant \quad &\mathbb{E}_\theta\left[\left|\int \hat{q}(\theta)f(\theta)d\theta - \int q(\theta)f(\theta)d\theta\right|\right] + \mathbb{E}_\theta\left[\left|\int \tilde{q}(\theta)f(\theta)d\theta - \int \hat{q}(\theta)f(\theta)d\theta\right|\right] \\
\leqslant \quad &\frac{2\sqrt{C}\|f(\theta)\|_\infty}{\sqrt{m}}
\end{aligned}$$

$\blacksquare$

With the knowledge of $p(\theta)$ and $q(\theta)$, we set $q_t(\theta) = \alpha_t(\theta)p(\theta)$, the PMD algorithm will reduce to adjust $\alpha(\theta_i)$ for samples $\{\theta_i\}_{i=1}^m \sim \pi(\theta)$ according to the stochastic gradient. Plug the gradient formula into the exact update rule, we have

$$q_{t+1}(\theta) = \frac{q_t(\theta)\exp(-\gamma_t g_t(\theta))}{Z} = \frac{\alpha_t(\theta)\exp(-\gamma_t g_t(\theta))p(\theta)}{Z} = \alpha_{t+1}(\theta)p(\theta)$$

where $\alpha_{t+1}(\theta) = \frac{\alpha_t(\theta)\exp(-\gamma_t g_t(\theta))}{Z}$. Since Z is constant, ignoring it will not effect the multiplicative update.

Given the fact that the objective function, $L(q)$, is 1-*strongly convex* w.r.t. the KL-divergence, we can immediately arrive at the following convergence results as appeared in Nemirovski et al. (2009), if we are able to compute the prox-mapping in Eq.(2) exactly.

Lemma 8 *One prox-mapping step Eq.(2) reduces the error by*

$$\mathbb{E}[KL(q^*||q_{t+1})] \leqslant (1-\gamma_t)\mathbb{E}[KL(q^*||q_t)] + \frac{\gamma_t^2 \mathbb{E}\|g_t\|_\infty^2}{2}.$$

With stepsize $\gamma_t = \frac{\eta}{t}$, it implies

$$\mathbb{E}[KL(q^*||q_T)] \leqslant \max\left\{KL(q^*||q_1), \frac{\eta^2\mathbb{E}\|g\|_\infty^2}{2\eta-1}\right\}\frac{1}{T}$$

Proof We could obtain the recursion directly from Theorem 2 by setting $\epsilon = 0$, which means solving the prox-mapping exactly, and the rate of convergence rate could be obtained by solving the recursion as stated in (Nemirovski et al., 2009). ■

Lemma 9 *Let q_t is the exact solution of the prox-mapping at t-step, then $\forall f(\theta) : \mathbb{R}^d \to \mathbb{R}$, which is bounded and integrable, we have*

$$\mathbb{E}\left[\left|\int q_t(\theta)f(\theta)d\theta - \int q(\theta)f(\theta)d\theta\right|\right] \leqslant \max\left\{\sqrt{KL(q^*||q_1)}, \frac{\eta\mathbb{E}\|g\|_\infty}{\sqrt{2\eta-1}}\right\}\frac{\|f\|_\infty}{\sqrt{t}}.$$

Proof

$$\begin{aligned}
&\mathbb{E}\left[\left|\int q_t(\theta)f(\theta)d\theta - \int q^*(\theta)f(\theta)d\theta\right|\right] = \mathbb{E}\|\langle q_t(\theta) - q^*(\theta)\, , f(\theta)\rangle_{L_2}\|_2 \\
&\leqslant \mathbb{E}[\|q_t(\theta) - q^*(\theta)\|_1\|f\|_\infty] \leqslant \|f\|_\infty\mathbb{E}[\|q_t(\theta) - q^*(\theta)\|_1] \leqslant \|f\|_\infty\mathbb{E}\left[\sqrt{\frac{1}{2}KL(q^*||q_t)}\right] \\
&\leqslant \max\left\{\sqrt{KL(q^*||q_1)}, \frac{\eta\mathbb{E}\|g\|_\infty}{\sqrt{2\eta-1}}\right\}\frac{\|f\|_\infty}{\sqrt{t}}
\end{aligned}$$

The second last inequality comes from Pinsker's inequality. ■

Theorem 5 *Assume the particle proposal prior $p(\theta)$ has the same support as the true posterior $q^*(\theta)$, i.e., $0 \leqslant q^*(\theta)/p(\theta) \leqslant C$. With further condition about the model $\|p(x|\theta)^N\|_\infty \leqslant \rho, \forall x$, then $\forall f(\theta) : \mathbb{R}^d \to \mathbb{R}$ bounded and integrable, with stepsize $\gamma_t = \frac{\eta}{t}$, the PMD algorithm return m weighted particles after T iteration such that*

$$\begin{aligned}
&\mathbb{E}\left[\left|\int \tilde{q}_t(\theta)f(\theta)d\theta - \int q^*(\theta)f(\theta)d\theta\right|\right] \\
\leqslant\ &\frac{2\sqrt{\max\{C, \rho\exp(\|g(\theta)\|_\infty)\}}\|f\|_\infty}{\sqrt{m}} + \max\left\{\sqrt{KL(q^*||\pi)}, \frac{\eta\mathbb{E}\|g\|_\infty}{\sqrt{2\eta-1}}\right\}\frac{\|f\|_\infty}{\sqrt{T}}.
\end{aligned}$$

Proof for Theorem 5.

We first decompose the error into optimization error and finite approximation error.

$$\begin{aligned}
&\mathbb{E}\left[\left|\int \tilde{q}_t(\theta)f(\theta)d\theta - \int q^*(\theta)f(\theta)d\theta\right|\right] \\
\leqslant\ &\underbrace{\mathbb{E}\left[\left|\int \tilde{q}_t(\theta)f(\theta)d\theta - \int q_t(\theta)f(\theta)d\theta\right|\right]}_{\text{finite approximation error } \epsilon_1} + \underbrace{\mathbb{E}\left[\left|\int q_t(\theta)f(\theta)d\theta - \int q^*(\theta)f(\theta)d\theta\right|\right]}_{\text{optimization error } \epsilon_2}
\end{aligned}$$

For the optimization error, by lemma 9, we have

$$\epsilon_2 \leqslant \max\left\{\sqrt{KL(q^*||q_1)}, \frac{\eta\mathbb{E}\|g\|_\infty}{\sqrt{2\eta-1}}\right\}\frac{\|f\|_\infty}{\sqrt{t}}.$$

Recall that

$$\begin{aligned} q_t(\theta) = & \frac{q_{t-1}(\theta)\exp(-\gamma_{t-1}g_{t-1}(\theta))}{Z} \\ = & \frac{\alpha_{t-1}(\theta)p(\theta)(\alpha_{t-1}^{-\gamma_{t-1}}(\theta)p(x|\theta)^{N\gamma_{t-1}})}{Z} = \alpha_{t-1}^{1-\gamma_{t-1}}(\theta)p(\theta)\frac{p(x|\theta)^{N\gamma_{t-1}}}{Z} \end{aligned}$$

which results the update $\alpha_t(\theta) = \frac{\alpha_{t-1}^{1-\gamma_{t-1}}(\theta)p(x|\theta)^{N\gamma_{t-1}}}{Z}$. Notice $Z = \int q_t(\theta)\exp(-\gamma_t g_t(\theta))d\theta$, we have $\exp(-\gamma_t\|g_t(\theta)\|_\infty) \leqslant Z \leqslant \exp(\gamma_t\|g_t(\theta)\|_\infty)$. By induction, it can be show that $\|\alpha_t\|_\infty \leqslant \max\{C, \rho\exp(\|g_t(\theta)\|_\infty)\} \leqslant \max\{C, \rho\exp(\|g(\theta)\|_\infty)\}$. Therefore, by lemma 7, we have

$$\epsilon_1 \leqslant \frac{2\sqrt{\max\{C, \rho\exp(\|g(\theta)\|_\infty)\}}\|f\|_\infty}{\sqrt{m}}.$$

Combine ϵ_1 and ϵ_2, we achieve the conclusion. ■

Remark: Simply induction without the assumption from the update of $\alpha_t(\theta)$ will result the upper bound of sequence $\|\alpha_t\|_\infty$ growing. The growth of sequence $\|\alpha_t\|_\infty$ is also observed in the proof (Crisan and Doucet, 2002) for sequential Monte Carlo on dynamic models. To achieve the uniform convergence rate for SMC of inference on dynamic system, Crisan and Doucet (2002); Gland and Oudjane (2004) require the models should satisfy i), $\epsilon\nu(\theta_i) \leqslant p(x_i|\theta_i)p(\theta_i|\theta_{i-1}) \leqslant \epsilon^{-1}\nu(\theta_i)$, $\forall x$ where $\nu(\theta)$ is a positive measure, and ii), $\frac{\sup_\theta p(x|\theta)}{\inf_{\mu\in\mathcal{P}}\langle\mu(\theta)p(\cdot|\theta)p(x|\cdot)\rangle} \leqslant \rho$. Such rate is only for SMC on dynamic system. For static model, the transition distribution is unknown, and therefore, no guarantee is provided yet. With much simpler and more generalized condition on the model, *i.e.*, $\|p(x|\theta)^N\|_\infty \leqslant \rho$, we also achieve the uniform convergence rate for static model. There are plenties of models satisfying such condition. We list several such models below.

1. logistic regression, $p(y|x,w) = \frac{1}{1+\exp(-yw^\top x)}$, and $\|p(y|x,w)\|_\infty \leqslant 1$.

2. probit regression, $p(y=1|x,w) = \Phi(w^\top x)$ where $\Phi(\cdot)$ is the cumulative distribution function of normal distribution. $\|p(y|x,w)\|_\infty \leqslant 1$.

3. multi-category logistic regression, $p(y=k|x,W) = \frac{\exp(w_k^\top x)}{\sum_{i=1}^K \exp(w_k^\top x)}$, and $\|p(y|x,W)\|_\infty \leqslant 1$.

4. latent Dirichlet allocation,

$$\begin{aligned} p(x_d|\theta_d,\Phi) &= \mathbb{E}_{z_d\sim p(z_d|\theta_d)}[p(x_d|z_d,\Phi)] \\ p(x_d|z_d,\Phi) &= \prod_{n=1}^{N_d}\prod_{w=1}^{W}\prod_{k=1}^{K}\Phi_{kw}^{z_{dnk}x_{dnw}} \\ p(z_d|\theta_d) &= \prod_{n=1}^{N_d}\prod_{k=1}^{K}\theta_{dk}^{z_{dnk}} \end{aligned}$$

   and $\|p(x_d|\theta_d,\Phi)\|_\infty \leqslant \max_{z_d}\|p(x_d|z_d,\Phi)\|_\infty \leqslant 1$.

5. linear regression, $p(y|w,x) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-(y-w^\top x)^2/2\sigma^2)$, and $\|p(y|w,x)\|_\infty \leqslant \frac{1}{\sigma\sqrt{2\pi}}$.

6. Gaussian model and PCA, $p(x|\mu,\Sigma) = (2\pi\det(\Sigma))^{-\frac{1}{d}}\exp\bigg(-\frac{1}{2}(x-\mu)^\top\Sigma(x-\mu)\bigg)$, and $\|p(x|\mu,\Sigma)\|_\infty \leqslant (2\pi\det(\Sigma))^{-\frac{1}{d}}$.

D Error Bound of Weighted Kernel Density Estimator

Before we start to prove the finite convergence in general case, we need to characterize the error induced by weighted kernel density estimator. In this section, we analyze the error in terms of both L_1 and L_2 norm, which are used for convergence analysis measured by KL-divergence in Appendix E .

D.1 L_1-Error Bound of Weighted Kernel Density Estimator

We approximate the density function $q(\theta)$ using the weighted kernel density estimator $\tilde{q}(\theta)$ and would like to bound the L_1 error, i.e. $\|\tilde{q}(\theta) - q(\theta)\|_1$ both in expectation and with high probability. We consider an unnormalized kernel density estimator as the intermediate quantity

$$\varrho_m(\theta) = \frac{1}{m}\sum_{i=1}^{m} \omega(\theta_i) K_h(\theta, \theta_i)$$

Note that $\mathbb{E}[\varrho_m(\theta)] = \mathbb{E}_{\theta_i}[\omega(\theta_i)K_h(\theta,\theta_i)] = q \star K_h$. Then the error can be decomposed into three terms as

$$\epsilon := \mathbb{E}\left\|\tilde{q}(\theta) - q(\theta)\right\|_1 \leqslant \underbrace{\mathbb{E}\left\|\tilde{q}(\theta) - \varrho_m(\theta)\right\|_1}_{\text{normalization error}} + \underbrace{\mathbb{E}\left\|\varrho_m(\theta) - \mathbb{E}\,\varrho_m(\theta)\right\|_1}_{\text{sampling error (variance)}} + \underbrace{\left\|\mathbb{E}\,\varrho_m(\theta) - q(\theta)\right\|_1}_{\text{approximation error (bias)}}$$

We now present the proof for each of these error bounds.

To formally show that, we begin by giving the definition of a special class of kernels and Hölder classes of densities that we consider.

Definition 10 (($\beta;\mu,\nu,\delta$)-valid density kernel) *We say a kernel function $K(\cdot)$ is a $(\beta;\mu,\nu)$-valid density kernel, if $K(\theta,\theta) = K(\theta-\theta)$ is a bounded kernel such that*

(i) $\int K(z)dz = 1$

(ii) $\int |K(z)|^r dz \leqslant \infty$ *for any* $r \geqslant 1$*, particularly,* $\int K(z)^2\, dz \leqslant \mu^2$ *for some* $\mu > 0$*.*

(iii) $\int z^s K(z)dz = 0$*, for any* $s = (s_1,\ldots,s_d) \in \mathbb{N}^d$ *such that* $1 \leqslant |s| \leqslant \lfloor\beta\rfloor$*. In addition,* $\int \|z\|^\beta |K(z)|dz \leqslant \nu$ *for some* $\nu > 0$*.*

For simplicity, we sometimes call $K(\cdot)$ as a β-valid density kernel if the constants μ and ν are not specifically given. Notice that all spherically symmetric probability density and product kernels based on symmetric univariate densities satisfy the conditions. For instance, the kernel $K(\theta) = (2\pi)^{-d/2}\exp(-\|\theta\|^2/2)$ satisfies the conditions with $\beta = 1$, and it is used through out our experiments. Furthermore, we will focus on a class of smooth densities

Definition 11 (($\beta;\mathcal{L}$)-Hölder density function) *We say a density function $q(\cdot)$ is a $(\beta;\mathcal{L})$-Hölder density function if function $q(\cdot)$ is $\lfloor\beta\rfloor$-times continuously differentiable on its support Ω and satisfies*

(i) for any z_0, there exists $L(z_0) > 0$ such that

$$|q(z) - q_{z_0}^{(\beta)}(z)| \leqslant L(z_0)\|z - z_0\|^\beta, \forall z \in \Omega$$

where $q_{z_0}^{(\beta)}$ is the $\lfloor\beta\rfloor$-order Taylor approximation, i.e.

$$q_{z_0}^{(\beta)}(z) := \sum_{s=(s_1,\ldots,s_d):|s|\leqslant\lfloor\beta\rfloor} \frac{(z-z_0)^s}{s!} D^s q(z_0);$$

(ii) in addition, the integral $\int L(z)dz \leqslant \mathcal{L}$.

$f \in C_{\mathcal{L}}^{\beta}(\Omega)$ means f is $(\beta;\mathcal{L})$-Hölder density function.

Then given the above setting for the kernel function and the smooth densities, we can characterize the error of the weighted kernel density estimator as follows.

D.1.1 KDE error due to bias

Lemma 12 (Bias) *If $q(\cdot) \in C_{\mathcal{L}}^{\beta}(\Omega)$ and K is a $(\beta;\mu,\nu)$-valid density kernel, then*

$$\left\|q(\theta) - \mathbb{E}[\varrho_m(\theta)]\right\|_1 \leqslant \nu\mathcal{L}h^\beta.$$

Proof The proof of this lemma follows directly from Chapter 4.3 in (Wand and Jones, 1995).

$$\begin{aligned} |\mathbb{E}[\varrho_m(\theta)] - q(\theta)| &= |q \star K_h(\theta) - q(\theta)| \\ &= \int \frac{1}{h^d} K(\frac{z-\theta}{h}) q(z) dz - q(\theta) \\ &= \int \frac{1}{h^d} K(\frac{z}{h})[q(\theta+z) - q(\theta)]dz \\ &= \int K(z)[q(\theta+hz) - q(\theta)]dz \\ &\leqslant \left|\int K(z)[q(\theta+hz) - q_\theta^{(\beta)}(\theta+hz)]dz\right| + \int \left|K(z)[q_\theta^{(\beta)}(\theta+hz) - q(\theta)]dz\right| \\ &\leqslant L(\theta)\int |K(z)|\|hz\|^\beta dz + \left|\int K(z)[q_\theta^{(\beta)}(\theta+hz) - q(\theta)]dz\right| \end{aligned}$$

Note that $q_\theta^{(\beta)}(\theta+hz) - q(\theta)$ is a polynomial of degree at most $\lfloor\beta\rfloor$ with no constant, by the definition of $(\beta;\mu,\nu)$-valid density kernel, the second term is zero. Hence, we have $|\mathbb{E}[\varrho_m(\theta)] - q(\theta)| \leqslant \nu L(\theta)h^\beta$, and therefore

$$\|\mathbb{E}[\varrho_m(\theta)] - q(\theta)\|_1 \leqslant \nu h^\beta \int L(\theta)d\theta \leqslant \nu\mathcal{L}h^\beta.$$

■

D.1.2 KDE error due to variance

The variance term can be bounded using similar techniques as in (Devroye and Györfi, 1985).

Lemma 13 (Variance) *Assume $\omega\sqrt{p} \in L_1$ with bounded support, then*

$$\mathbb{E}\,\|\varrho_m(\theta) - \mathbb{E}[\varrho_m(\theta)]\|_1 \leqslant \frac{\mu}{\sqrt{m}h^{\frac{d}{2}}}\int \omega\sqrt{p}\,d\theta + o((mh^d)^{-\frac{1}{2}}).$$

Proof For any θ, we have

$$\begin{aligned} \sigma^2(\theta) :&= \mathbb{E}\left[(\varrho_m(\theta) - \mathbb{E}[\varrho_m(\theta)])^2\right] \\ &= \frac{1}{m}\sum_{i=1}^m \mathbb{E}[\omega^2(\theta_i)K_h^2(\theta,\theta_i)] - (q \star K_h)^2 \leqslant \frac{(\omega^2 q) \star K_h^2}{m} \end{aligned}$$

Denote $\mu(K) := \sqrt{\int K(\theta)^2\,d\theta}$ and kernel $K^+(\theta) = \frac{K^2(\theta)}{\mu(K)^2}$, then $\mu(K) \leqslant \mu$, $\int K^+ d\theta = 1$ and

$$K_h^+(\theta) = \frac{1}{h^d}K^+(\theta/d) = \frac{1}{h^d}\frac{K(\theta/h)K(\theta/h)}{\mu^2(K)} = \frac{h^d}{\mu^2(K)}K_h^2(\theta)$$

Hence,

$$\sigma^2(\theta) \leqslant \frac{\mu^2(K)(\omega^2 p) \star K_h^+}{mh^d} \leqslant \frac{\mu^2[(\omega^2 p) \star K_h^+ - \omega^2 p]}{mh^d} + \frac{\mu^2(\omega^2 p)}{mh^d}.$$

Note that $\sigma(\theta) = \sqrt{\mathbb{E}\left[(\varrho_m(\theta) - \mathbb{E}[\varrho_m(\theta)])^2\right]} \geqslant \mathbb{E}|\varrho_m(\theta) - \mathbb{E}[\varrho_m(\theta)]|$, hence

$$\begin{aligned} &\mathbb{E}\,\|\varrho_m(\theta) - \mathbb{E}[\varrho_m(\theta)]\|_1 \\ &= \int \mathbb{E}|\varrho_m(\theta) - \mathbb{E}[\varrho_m(\theta)]|\,d\theta \leqslant \int \sigma(\theta)\,d\theta \\ &\leqslant \int \sqrt{\frac{\mu^2(\omega^2 p) \star K_h^+ - \omega^2 p]}{mh^d}} + \sqrt{\frac{\mu^2(\omega^2 p)}{mh^d}}\,d\theta \\ &\leqslant \frac{\mu}{\sqrt{m}h^{d/2}}\left[\int \sqrt{\omega^2 p}\,d\theta + \int \sqrt{(\omega^2 p) \star K_h^+ - \omega^2 p}\,d\theta\right] \\ &\leqslant \frac{\mu}{\sqrt{m}h^{d/2}}\left[\int \omega\sqrt{p}\,d\theta + \sqrt{|\Omega|}\cdot\sqrt{\int \left|(\omega^2 p) \star K_h^+ - \omega^2 p\right|\,d\theta}\right] \end{aligned}$$

From Theorem 2.1 in (Devroye and Györfi, 1985), we have $\int \left|(\omega^2 p) \star K_h^+ - \omega^2 p\right|\,d\theta = o(1)$. Therefore, we conclude that

$$\mathbb{E}\left\|\varrho_m(\theta)-\mathbb{E}[\varrho_m(\theta)]\right\|_1 \leqslant \frac{\mu}{\sqrt{m}h^{d/2}}\|\omega\sqrt{p}\|_1 + o((mh^d)^{-\frac{1}{2}}).$$

■

D.1.3 KDE error due to normalization

The normalization error term can be easily derived based on the variance.

Lemma 14 (Normalization error) *Assume* $\omega\sqrt{p} \in L_2$

$$\mathbb{E}\left\|\tilde{q}(\theta)-\varrho_m(\theta)\right\|_1 \leqslant \frac{1}{\sqrt{m}}\left(\int \omega^2(\theta)p(\theta)\,d\theta\right)^{1/2}.$$

Proof Denote $\omega_i := \omega(\theta_i)$, then $\mathbb{E}[\omega_i] = \int \omega(\theta)p(\theta)\,d\theta = 1$ and $\mathbb{E}[\omega_i^2] = \int \omega^2(\theta)p(\theta)\,d\theta$, for any $i = 1,\ldots,m$. Hence,

$$\mathbb{E}|\frac{1}{m}\sum_{i=1}^m \omega_i - 1|^2 = \frac{1}{m}\int \omega^2(\theta)p(\theta)\,d\theta.$$

Recall that $\tilde{q}(\theta) = \frac{1}{\sum_{i=1}^m \omega_i}\sum_{i=1}^m \omega_i K_h(\theta,\theta_i)$ and $\varrho_m(\theta) = \frac{1}{m}\sum_{i=1}^m \omega_i K_h(\theta,\theta_i)$.

$$\begin{aligned}
&\mathbb{E}\left\|\tilde{q}(\theta)-\varrho_m(\theta)\right\|_1 \\
\leqslant\ &\mathbb{E}\left\|\frac{1}{\sum_{i=1}^m \omega_i}\sum_{i=1}^m \omega_i K_h(\theta,\theta_i) - \frac{1}{m}\sum_{i=1}^m \omega_i K_h(\theta,\theta_i)\right\|_1 \\
\leqslant\ &\mathbb{E}\left\|\left|1-\frac{\sum_{i=1}^m \omega_i}{m}\right|\frac{1}{\sum_{i=1}^m \omega_i}\sum_{i=1}^m \omega_i K_h(\theta,\theta_i)\right\|_1 \\
\leqslant\ &\mathbb{E}\left|1-\frac{\sum_{i=1}^m \omega_i}{m}\right|\cdot\|K_h(\theta)\|_1
\end{aligned}$$

Since $\|K_h\|_1 = \int \frac{1}{h^d}K(\theta/h)\,d\theta = \int K(\theta)\,d\theta = 1$, we have

$$\mathbb{E}\left\|\tilde{q}(\theta)-\varrho_m(\theta)\right\|_1 \leqslant \frac{1}{\sqrt{m}}\sqrt{\int \omega^2(\theta)p(\theta)\,d\theta} = \frac{1}{\sqrt{m}}\|w\sqrt{p}\|_2$$

■

D.1.4 KDE error in expectation and with high probability

Based on the above there lemmas, namely, Lemma 12 - 14, we can immediately arrive at the bound of the L_1 error in expectation as stated in Theorem 4. We now provide the proof for the high probability bound as stated below.

Corollary 15 (Overall error in high probability) *Besides the above assumption, let us also assume that* $\omega(\theta)$ *is bounded, i.e. there exists* $0 < B_1 \leqslant B_2 < \infty$ *such that* $B_1 \leqslant \omega(\theta) \leqslant B_2, \forall\theta$*. Then, with probability at least* $1-\delta$*,*

$$\|\tilde{q}(\theta)-q(\theta)\|_1 \leqslant \nu\mathcal{L}h^\beta + \frac{\mu}{\sqrt{m}h^{d/2}}\|\omega\sqrt{p}\|_1 + \frac{1}{\sqrt{m}}\|\omega\sqrt{p}\|_2 + \frac{1}{\sqrt{m}}\sqrt{8B_1B_2\log(1/\delta)} + o((mh^d)^{-\frac{1}{2}}).$$

Proof We use McDiarmid's inequality to show that the function $f(\Theta) = \left\|\tilde{q}(\theta)-q(\theta)\right\|_1$, defined on the random data $\Theta = (\theta_1,\ldots,\theta_m)$, is concentrated on the mean. Let $\tilde{\Theta} = (\theta_1,\ldots,\tilde{\theta}_j,\ldots,\theta_m)$. We denote $\omega = (\omega(\theta_1),\ldots,\omega(\theta_m))$ and $\tilde{\omega} = (\omega(\theta_1),\ldots,\omega(\tilde{\theta}_j),\ldots,\omega(\theta_m))$. Denote $k = (K_h(\theta,\theta_1),\ldots,K_h(\theta,\theta_m))$ and $\tilde{k} =$

$(K_h(\theta,\theta_1),\ldots,K_h(\theta,\theta_j'),\ldots,K_h(\theta,\theta_m))$. We first show that $|f(\Theta)-f(\Theta')|$ is bounded.

$$\begin{aligned}
& |f(\Theta)-f(\Theta')| \\
= \; & \left|\, \|\tilde{q}_{\Theta}(\theta)-q(\theta)\|_1 - \|\tilde{q}_{\tilde{\Theta}}(\theta)-q(\theta)\|_1 \,\right| \\
\leqslant \; & \|\tilde{q}_{\Theta}(\theta)-\tilde{q}_{\tilde{\Theta}}(\theta)\|_1 \\
= \; & \left\| \frac{\sum_{i=1}^m \omega_i k_i}{\sum_{i=1}^m \omega_i} - \frac{\sum_{i=1}^m \tilde{\omega}_i \tilde{k}_i}{\sum_{i=1}^m \tilde{\omega}_i} \right\|_1 \\
\leqslant \; & \left\| \frac{(\tilde{\omega}_j-\omega_j)\cdot(\sum_{i=1}^m \omega_i k_i) - (\sum_{i=1}^m \omega_i)(\tilde{\omega}_i\tilde{k}_i - \omega_j k_j)}{(\sum_{i=1}^m \omega_i)\cdot(\sum_{i=1}^m \tilde{\omega}_i)} \right\|_1 \\
\leqslant \; & \left\| \frac{\tilde{\omega}_j-\omega_j}{(\sum_{i=1}^m \tilde{\omega}_i)} \right\|_\infty + \left\| \frac{\tilde{\omega}_i\tilde{k}_i - \omega_j k_j}{\sum_{i=1}^m \tilde{\omega}_i} \right\|_1 \\
\leqslant \; & \frac{2B_1B_2}{m} + \frac{2B_1B_2}{m} \leqslant \frac{4B_1B_2}{m}
\end{aligned}$$

Invoking the McDiamid's inequality, we have

$$\Pr\left(f(\Theta) - \mathbb{E}_\Theta[f(\Theta)] \geqslant \epsilon\right) \leqslant \exp\left\{-\frac{m\epsilon^2}{8B_1^2B_2^2}\right\}, \forall \epsilon > 0$$

which implies the corollary.

■

D.2 L_2-Error Bound of Weighted Kernel Density Estimator

Following same argument yields also similar L_2-error bound of the weighted kernel density estimator, i.e. $\|\tilde{q}(\theta)-q(\theta)\|_2$. For completeness and also for future reference, we provide the exact statement of the bound below in line with Theorem 4 and Corollary 15.

Theorem 16 (L_2-error in expectation) *Let $q=\omega p \in C_{\mathcal{L}}^{\beta}(\Omega)$ and K be a $(\beta;\mu,\nu)$-valid density kernel. Assume that $\omega^2 p \in L_2$ and has bounded support. Then*

$$\mathbb{E}\,\|\tilde{q}(\theta)-q(\theta)\|_2^2 \leqslant 2(\nu h^\beta \mathcal{L})^2 + \frac{8\mu^2}{mh^d}\|\omega\sqrt{p}\|_2^2 + o((mh^d)^{-1}).$$

Proof for Theorem 16. The square L_2-error can also be decomposed into three terms.

$$\mathbb{E}\,\|\tilde{q}(\theta)-q(\theta)\|_2^2 \leqslant 4\underbrace{\mathbb{E}\,\|\tilde{q}(\theta)-\varrho_m(\theta)\|_2}_{\text{normalization error}} + 4\underbrace{\mathbb{E}\,\|\varrho_m(\theta)-\mathbb{E}\,\varrho_m(\theta)\|_2^2}_{\text{sampling error (variance)}} + 2\underbrace{\|\mathbb{E}\,\varrho_m(\theta)-q(\theta)\|_2^2}_{\text{approximation error (bias)}}$$

This uses the inequality $(a+b+c)^2 \leqslant 2a^2+4b^2+4c^2$ for any a,b,c. From Lemma 12, we already have $|\mathbb{E}[\varrho_m(\theta)]-q(\theta)| \leqslant L(\theta)\int |K(z)|\|hz\|^\beta dz, \forall\theta$. Hence,

$$\|\mathbb{E}[\varrho_m(\theta)]-q(\theta)\|_2^2 \leqslant \nu^2 h^{2\beta}\int L^2(\theta)d\theta \leqslant (\nu h^\beta \mathcal{L})^2. \tag{12}$$

From proof for Lemma 13, we have

$$\mathbb{E}\,\|\varrho_m(\theta)-\mathbb{E}[\varrho_m(\theta)]\|_2^2 = \textstyle\int \mathbb{E}|\varrho_m(\theta)-\mathbb{E}[\varrho_m(\theta)]|^2\,d\theta \leqslant \int \sigma^2(\theta)\,d\theta \tag{13}$$

$$\leqslant \textstyle\int \frac{\mu^2[(\omega^2 p)\star K_h^+ - \omega^2 p]}{mh^d} + \frac{\mu^2(\omega^2 p)}{mh^d}\,d\theta \leqslant \frac{\mu^2}{mh^d}\|\omega\sqrt{p}\|_2^2 + o((mh^d)^{-1}) \tag{14}$$

In addition, we have for the normalization error term,

$$\mathbb{E}\,\|\tilde{q}(\theta)-\varrho_m(\theta)\|_2^2 \leqslant \mathbb{E}\left\|\left(1-\tfrac{\sum_{i=1}^m \omega_i}{m}\right)\tfrac{\sum_{i=1}^m \omega_i K_h(\theta,\theta_i)}{\sum_{i=1}^m \omega_i}\right\|_2^2 \tag{15}$$

$$\leqslant \mathbb{E}\left|1-\tfrac{\sum_{i=1}^m \omega_i}{m}\right|^2 \cdot \|K_h\|_2^2 \leqslant \tfrac{\mu^2}{mh^d}\|\omega\sqrt{p}\|_2^2$$

Combining equation (12) , (13) and (15), it follows that

$$\mathbb{E}\,\|\tilde{q}(\theta)-q(\theta)\|_2^2 \leqslant 2(\nu h^\beta \mathcal{L})^2 + \frac{8\mu^2}{mh^d}\|\omega\sqrt{p}\|_2^2 + o((mh^d)^{-1}).$$

■

Corollary 17 (**L_2-error in high probability**) *Besides the above assumption, let us also assume that $\omega(\theta)$ is bounded, i.e. there exists $0 < B_1 \leqslant B_2 < \infty$ such that $B_1 \leqslant \omega(\theta) \leqslant B_2, \forall \theta$. Then, with probability at least $1 - \delta$,*

$$\|\tilde{q}(\theta) - q(\theta)\|_2^2 \leqslant 2(\nu h^\beta \mathcal{L})^2 + \frac{8\mu^2}{mh^d}\|\omega\sqrt{p}\|_2^2 + o((mh^d)^{-1}) + \frac{16 B_1 B_2 \mu^2}{m}\sqrt{\log(1/\delta)}.$$

Proof for Theorem 17. Use McDiarmid's inequality similar as proof for Corollary 15. ■

E Convergence Analysis for Density Approximation

In this section, we consider the rate of convergence for the entire density measured by KL-divergence. We start with the following lemma that show the renormalization does not effect the optimization in the sense of optimal, and we show the importance weight $\omega_t(\theta) = \frac{\exp(-\gamma_t g_t(\theta))}{Z}$ at each step are bounded under proper assumptions. Moreover, the error of the prox-mapping at each step incurred by the weighted density kernel density estimation is bounded.

Lemma 18 *Let $\zeta = \int_{\backslash\Omega} \tilde{q}_t d\theta$, $\widehat{q}_t = \frac{\tilde{q}_t}{1-\zeta}$ is a valid density on Ω, then, $\tilde{q}_t^+ = \widehat{q}_t^+$, where $\tilde{q}_t^+ := \operatorname{argmin}_{q \in \mathcal{P}(\Omega)} F_t(q; \tilde{q}_t)$, $\widehat{q}_t^+ := \operatorname{argmin}_{q \in \mathcal{P}(\Omega)} F_t(q; \widehat{q}_t)$, and $F_t(q; q') := \langle q, \gamma_t g \rangle_{L_2} + KL(q\|q')$.*

Proof for Lemma 18. The minima of prox-mapping is not effected by the renormalization. Indeed, such fact can be verified by comparing to $\tilde{q}_t^+ = \operatorname{argmin} F_t(q; \tilde{q}_t)$ and $\widehat{q}_t^+ = \operatorname{argmin} F_t(q; \widehat{q}_t)$, respectively.

$$\widehat{q}_t^+ = \frac{(\frac{1}{1-\zeta}\tilde{q}_t)^{1-\gamma_t} p(\theta)_t^\gamma p(x_t|\theta)^{N\gamma_t}}{\int (\frac{1}{1-\zeta}\tilde{q}_t)^{1-\gamma_t} p(\theta)_t^\gamma p(x_t|\theta)^{N\gamma_t} d\theta} = \frac{\tilde{q}_t^{1-\gamma_t} p(\theta)_t^\gamma p(x_t|\theta)^{N\gamma_t}}{\int \tilde{q}_t^{1-\gamma_t} p(\theta)_t^\gamma p(x_t|\theta)^{N\gamma_t} d\theta} = \tilde{q}_t^+$$

■

Due to the fact, we use $\tilde{q}_t^+$ following for consistency. Although the algorithm updates based on $\tilde{q}_t$, it is implicitly doing renoramlization after each update. We will show that $\widehat{q}_{t+1}$ is an ϵ-inexact prox-mapping.

Lemma 19 *Assume for all mini-batch of examples $\|g_t(\theta)\|_\infty^2 \leqslant M^2$, then we have*

(a) $\exp(-2\gamma_t M) \leqslant \omega_t(\theta) = \frac{\tilde{q}_t^+(\theta)}{\widehat{q}_t(\theta)} \leqslant \exp(2\gamma_t M)$,

(b) $\|\nabla F_t(\tilde{q}_t^+; \widehat{q}_t)\|_\infty \leqslant 3\gamma_t M$.

Proof for Lemma 19. Let $Z := \int q_t(\theta)\exp(-\gamma_t g_t(\theta))d\theta$. We have $\exp(-\gamma_t M) \leqslant Z \leqslant \exp(\gamma_t M)$.
(a) Since $\|g_t(\theta)\|_\infty^2 \leqslant M^2$, we have

$$\exp(-2\gamma_t M) \leqslant \omega_t(\theta) = \frac{\tilde{q}_t^+(\theta)}{\widehat{q}_t(\theta)} = \frac{\exp(-\gamma_t g_t(\theta))}{Z} \leqslant \exp(2\gamma_t M).$$

(b) Also, because $\nabla F_t(q_t^+) = \gamma_t g_t + \log\frac{\tilde{q}_t^+}{\widehat{q}_t} = \gamma_t g_t + \log(\omega_t)$, it immediately follows

$$\|\nabla F_t(\tilde{q}_t^+; \widehat{q}_t)\|_\infty \quad = \quad \|\gamma_t g_t + \log(\omega_t)\|_\infty \leqslant \gamma_t\|g_t\|_\infty + \|\log(\omega_t)\|_\infty \leqslant \gamma_t M + (2\gamma_t M) = 3\gamma_t M.$$

■

Lemma 20 *Let $\epsilon_t := F_t(\widehat{q}_{t+1}; \widehat{q}_t) - F_t(\tilde{q}_t^+; \widehat{q}_t)$, which implies $\widehat{q}_{t+1} \in P_{\widehat{q}_t}^{\epsilon_t}(\gamma_t g_t)$. Let the bandwidth at step t satisfies*

$$h_t = O(1) m_t^{-1/(d+2\beta)},$$

one can guarantee that

$$\mathbb{E}_\theta[\epsilon_t | x_{[t-1]}, \theta_{[t-1]}] \leqslant \quad O(1)(\mu^2 + \nu^2\mathcal{L}^2)\mu^2\Delta m_t^{-\frac{2\beta}{d+2\beta}} + O(1)M(\mu + \nu\mathcal{L})\gamma_t m_t^{-\frac{\beta}{d+2\beta}}$$

In addition, with probability at least $1 - 2\delta$ in $\theta_t | x_{[t-1]}, \theta_{[t-1]}$, we have

$$\epsilon_t \leqslant O(1)(\mu^2\sqrt{\log(1/\delta)} + \nu^2\mathcal{L}^2)\mu^2\Delta m_t^{-\frac{2\beta}{d+2\beta}} + O(1)M(\mu + \nu\mathcal{L} + \sqrt{\log(1/\delta)})\gamma_t m_t^{-\frac{\beta}{d+2\beta}}$$

where $O(1)$ is some constant.

Proof for Lemma 20.

Note that since $\tilde{q}_t^+(\theta) = \tilde{q}_t(\theta)\exp(-\gamma_t g_t(\theta))/Z$, where $\tilde{q}_t(\theta) = \sum_{i=1}^{m_t}\alpha_i K_{h_t}(\theta-\theta_i)$, and $g_t(\theta) = \log(\tilde{q}_t) - \log(p) - N\log(p(x_t|\theta))$. By our assumption, we have $\tilde{q}_t \in C_{\mathcal{L}}^{\beta}(\Omega)$ and $\exp(-\gamma_t g_t) \in C_{\mathcal{L}}^{\beta}(\Omega)$; hence, $\tilde{q}_t^+ \in C_{\mathcal{L}}^{\beta}(\Omega)$. Invoking the definition of function $F_t(\cdot;\widehat{q}_t)$, we have

$$\begin{aligned}
F_t(\widehat{q}_{t+1};\widehat{q}_t) - F_t(\tilde{q}_t^+;\widehat{q}_t) &= KL(\widehat{q}_{t+1}||\tilde{q}_t^+) + \langle\nabla F_t(\tilde{q}_t^+;\widehat{q}_t), \widehat{q}_{t+1} - \tilde{q}_t^+\rangle_{L_2}\\
&\leqslant KL(\widehat{q}_{t+1}||\tilde{q}_t^+) + 3\gamma_t M\|\tilde{q}_t^+ - \widehat{q}_{t+1}\|_1\\
&\leqslant \int\frac{(\widehat{q}_{t+1}-\tilde{q}_t^+)^2}{\tilde{q}_t^+}d\theta + 3\gamma_t M\|\tilde{q}_t^+ - \widehat{q}_{t+1}\|_1\\
&\leqslant \Delta\|\widehat{q}_{t+1} - \tilde{q}_t^+\|_2^2 + 3\gamma_t M\|\tilde{q}_t^+ - \widehat{q}_{t+1}\|_1
\end{aligned}$$

Based on the definition of $\widehat{q}_{t+1}$, we have

$$\begin{aligned}
\|\tilde{q}_t^+ - \widehat{q}_{t+1}\|_1 &= \left\|\frac{1}{1-\zeta}\tilde{q}_{t+1} - \tilde{q}_t^+\right\|_1 = \frac{1}{1-\zeta}\|\tilde{q}_{t+1} - \tilde{q}_t^+ + \zeta\tilde{q}_t^+\|_1 \leqslant \frac{1}{1-\zeta}\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_1 + \frac{\zeta}{1-\zeta}\\
&= \|\tilde{q}_{t+1} - \tilde{q}_t^+\|_1 + \zeta + o(\zeta + \|\tilde{q}_{t+1} - \tilde{q}_t^+\|_1).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|\tilde{q}_t^+ - \widehat{q}_{t+1}\|_2^2 &= \left\|\frac{1}{1-\zeta}(\tilde{q}_{t+1} - \tilde{q}_t^+) + \frac{\zeta}{1-\zeta}\tilde{q}_t^+\right\|_2^2\\
&\leqslant \frac{2}{(1-\zeta)^2}\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_2^2 + \frac{2\zeta^2}{(1-\zeta)^2}\|\tilde{q}_t^+\|_2^2\\
&\leqslant 2(1+\zeta)^2\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_2^2 + 2\zeta^2\|\tilde{q}_t^+\|_2^2 + o(\zeta^2\|\tilde{q}_t^+\|_2^2 + \zeta^2\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_2^2)
\end{aligned}$$

Recall $\zeta = 1 - \int_\Omega \tilde{q}_{t+1} = \langle 1, \tilde{q}_t^+ - \tilde{q}_{t+1}\rangle \leqslant \|\tilde{q}_{t+1} - \tilde{q}_t^+\|_1$, we can simplify the L_1 and L_2 error as

$$\begin{aligned}
\|\widehat{q}_{t+1} - \tilde{q}_t^+\|_1 &= 2\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_1 + o(\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_1),\\
\|\tilde{q}_t^+ - \widehat{q}_{t+1}\|_2^2 &\leqslant 2\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_2^2 + 2\|\tilde{q}_t^+\|_2^2\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_1^2 + o(\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_2^2 + \|\tilde{q}_t^+\|_2^2\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_1^2)\\
&\leqslant (2 + 2\|\tilde{q}_t^+\|_2^2)\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_2^2 + o(\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_2^2).
\end{aligned}$$

The last inequality for L_2 error comes from Jensen's inequality. We argue that $\|\tilde{q}_t^+\|_2^2$ is finite. Indeed,

$$\begin{aligned}
\|\tilde{q}_t^+\|_2^2 &= \int(\tilde{q}_t^+)^2 d\theta = \int\frac{\tilde{q}_t^2\exp(-2\gamma_t g_t)}{Z^2}d\theta \leqslant \left\|\frac{\exp(-2\gamma_t g_t)}{Z^2}\right\|_\infty\int\tilde{q}_t^2 d\theta\\
&\leqslant \exp(4\gamma_t M)\left(\sum_{i,j}\alpha_i^t\alpha_j^t\int K_h(\theta-\theta_i)K_h(\theta-\theta_j)d\theta\right)\\
&\leqslant \exp(4\gamma_t M)\left(\sum_{i,j}\alpha_i^t\alpha_j^t\|K_h(\theta-\theta_i)\|_2\|K_h(\theta-\theta_j)\|_2\right) \leqslant \exp(4\gamma_t M)\mu^2\|\alpha^t\|_1\|\alpha^t\|_\infty \leqslant \exp(4\gamma_t M)\mu^2
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\epsilon_t \leqslant\ & (2\Delta + 2\Delta\mu^2\exp(4\gamma_t M))\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_2^2 + 6\gamma_t M\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_1\\
& + o(\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_2^2 + \gamma_t\|\tilde{q}_{t+1} - \tilde{q}_t^+\|_1)
\end{aligned}$$

Applying the result of Theorem 4 and 16 for $\widehat{q}_{t+1}$ and $\tilde{q}_t^+$ we have

$$\begin{aligned}
\mathbb{E}_\theta[\epsilon_t|x_{[t-1]},\theta_{[t-1]}] \leqslant\ & (2\Delta + 2\Delta\mu^2\exp(4\gamma_t M))\left[2(\nu h_t^\beta\mathcal{L})^2 + \frac{8\mu^2}{m_t h_t^d}\|\omega_t\sqrt{\tilde{q}_t}\|_2^2 + o((m_t h_t^d)^{-1})\right]\\
& + 6\gamma_t M\left[\nu\mathcal{L}h_t^\beta + \frac{\mu}{\sqrt{m_t}h_t^{d/2}}\|\omega_t\sqrt{\tilde{q}_t}\|_1 + \frac{1}{\sqrt{m_t}}\|\omega_t\sqrt{\tilde{q}_t}\|_2 + o((m_t h_t^d)^{-\frac{1}{2}})\right]\\
& + o\left(2(\nu h_t^\beta\mathcal{L})^2 + \frac{8\mu^2}{m_t h_t^d}\|\omega_t\sqrt{\tilde{q}_t}\|_2^2 + \gamma_t[\nu\mathcal{L}h_t^\beta + \frac{\mu}{\sqrt{m_t}h_t^{d/2}}\|\omega_t\sqrt{\tilde{q}_t}\|_1 + \frac{1}{\sqrt{m_t}}\|\omega_t\sqrt{\tilde{q}_t}\|_2]\right)
\end{aligned}$$

Under the Assumption C, we already proved that $|\omega_t|_\infty \leqslant \exp(2\gamma_t M)$, hence, $\|\omega_t\sqrt{\tilde{q}_t}\|_2^2 \leqslant \exp(4\gamma_t M)$. Without loss of generality, we can assume $\int\sqrt{\tilde{q}_t(\theta)}d\theta \leqslant O(1)$ and $\gamma_t M \leqslant O(1)$ for all t, then we can simply write $\|\omega_t\sqrt{\tilde{q}_t}\|_1 \leqslant O(1)$

and $\|\omega_t\sqrt{\tilde{q}_t}\|_2^2 \leqslant O(1)$. When $h_t = O(1)m_t^{-1/(d+2\beta)}$, the above result can be simplified as

$$\mathbb{E}_\theta[\epsilon_t|x_{[t-1]},\theta_{[t-1]}] \leqslant \quad O(1)(\mu^2+\nu^2\mathcal{L}^2)\mu^2\Delta m_t^{-\frac{2\beta}{d+2\beta}} + O(1)M(\mu+\nu\mathcal{L})\gamma_t m_t^{-\frac{\beta}{d+2\beta}}$$

Similarly, combining the results of Corollary 15 and 17, we have with probability at least $1-2\delta$,

$$\begin{aligned}\epsilon_t \leqslant \quad & (2\Delta + 2\Delta\mu^2\exp(4\gamma_t M))\left[2(\nu h_t^\beta\mathcal{L})^2 + \frac{8\mu^2}{m_t h_t^d}\|\omega_t\sqrt{\tilde{q}_t}\|_2^2 + o((m_t h_t^d)^{-1}) + \frac{16B_1B_2\mu^2}{m_t}\sqrt{\log(1/\delta)}\right] \\ & + 6\gamma_t M\left[\nu\mathcal{L}h^\beta + \frac{\mu}{\sqrt{m_t}h_t^{d/2}}\|\omega_t\sqrt{\tilde{q}_t}\|_1 + \frac{1}{\sqrt{m_t}}\|\omega_t\sqrt{\tilde{q}_t}\|_2 + \frac{1}{\sqrt{m_t}}\sqrt{8B_1B_2\log(1/\delta)} + o((m_t h_t^d)^{-\frac{1}{2}})\right] \\ & + o\left(2(\nu h_t^\beta\mathcal{L})^2 + \frac{8\mu^2}{m_t h_t^d}\|\omega_t\sqrt{\tilde{q}_t}\|_2^2 + \gamma_t[\nu\mathcal{L}h_t^\beta + \frac{\mu}{\sqrt{m_t}h_t^{d/2}}\|\omega_t\sqrt{\tilde{q}_t}\|_1 + \frac{1}{\sqrt{m_t}}\|\omega_t\sqrt{\tilde{q}_t}\|_2]\right)\end{aligned}$$

which leads to the lemma.

■

Our main Theorem 6 follows immediately by applying the results in the above lemma to Theorem 2.

Proof of Theorem 6. We first notice that

$$\begin{aligned}\mathbb{E}[KL(q^*||\tilde{q}_T)] &= \mathbb{E}\left[\int q^*\log\frac{q^*}{\tilde{q}_T}d\theta\right] = \mathbb{E}\left[\int q^*\log\frac{q^*}{\widehat{q}_T}d\theta + \int q^*\log\frac{\widehat{q}_T}{\tilde{q}_T}d\theta\right] \\ &= \mathbb{E}[KL(q^*||\widehat{q}_T)] + \mathbb{E}\left[\int q^*\log\frac{\widehat{q}_T}{\tilde{q}_T}d\theta\right].\end{aligned}$$

For the second term,

$$\begin{aligned}\mathbb{E}\left[\int q^*\log\frac{\widehat{q}_T}{\tilde{q}_T}d\theta\right] &= \mathbb{E}\left[\langle q^*, \log\frac{\frac{1}{1-\zeta_T}\tilde{q}_T}{\tilde{q}_T}\rangle\right] = \mathbb{E}\left[\langle q^*, -\log(1-\zeta_T)\right] \\ &= \mathbb{E}[-\log(1-\zeta_T)] \leqslant \zeta_T + o(\zeta_T) \leqslant \mathbb{E}\|\tilde{q}_T - \tilde{q}_{T-1}^+\|_1 + o(\mathbb{E}\|\tilde{q}_T - \tilde{q}_{T-1}^+\|_1)\end{aligned}$$

By Theorem 4 and setting $h_t = O(1)m_t^{-1/(d+2\beta)}$, we achieve the error bound

$$\mathbb{E}\left[\int q^*\log\frac{\widehat{q}_T}{\tilde{q}_T}d\theta\right] \leqslant \mathcal{C}_2 m_t^{-\frac{\beta}{d+2\beta}},$$

where $\mathcal{C}_2 := O(1)M(\mu+\nu\mathcal{L})$.

When setting $\gamma_t = \min\{\frac{2}{t+1}, \frac{\Delta}{Mm_t^{\beta/(d+2\beta)}}\}$ invoking the above lemma, we have

$$\mathbb{E}_\theta[\epsilon_t|x_{[t-1]},\theta_{[t-1]}] \leqslant \mathcal{C}_1 m_t^{-2\beta/(d+2\beta)},$$

where $\mathcal{C}_1 := O(1)(\mu+\nu\mathcal{L})^2\mu^2\Delta$. Expanding the result from Theorem 2, it follows that

$$\mathbb{E}_{x,\theta}[KL(q^*||\widehat{q}_{t+1})] \leqslant (1-\gamma_t)\mathbb{E}_{x,\theta}[KL(q^*||\widehat{q}_t)] + \mathcal{C}_1 m_t^{-2\beta/(d+2\beta)} + \frac{\gamma_t^2}{2}M^2$$

The above recursion leads to the convergence result for the second term,

$$\mathbb{E}[KL(q^*||\widehat{q}_T)] \quad \leqslant \quad \frac{2\max\left\{D_1, M^2\right\}}{T} + \mathcal{C}_1\frac{\sum_{t=1}^T t^2 m_t^{-\frac{2\beta}{d+2\beta}}}{T^2}.$$

Combine these two results, we achieve the desired result

$$\mathbb{E}[KL(q^*||\tilde{q}_T)] \quad \leqslant \quad \frac{2\max\left\{D_1, M^2\right\}}{T} + \mathcal{C}_1\frac{\sum_{t=1}^T t^2 m_t^{-\frac{2\beta}{d+2\beta}}}{T^2} + \mathcal{C}_2 m_t^{-\frac{\beta}{d+2\beta}}.$$

■

Remark. The convergence in terms of KL-divergence is measuring the entire density and much more stringent compared to integral approximation. For the last iterate, an overall $O(\frac{1}{T})$ convergence rate can be achieved when $m_t = O(t^{2+d/\beta})$. Similar to Lemma 9, with Pinsker's inequality, we could easily obtain the the rate of convergence in terms of integral approximation from Theorem 6. After T steps, in general cases, the PMD algorithm converges in terms of integral approximation in rate $O(1/\sqrt{T})$ by choosing $O(1/t)$-decaying stepsizes and $O(t^{2+\frac{d}{2\beta}})$-growing samples.

F Derivation Details for Sparse Gaussian Processes and Latent Dirichlet Allocation

We apply the Particle Mirror Descent algorithm to sparse Gaussian processes and latent Dirichlet allocation. For these two models, we decompose the latent variables and incorporate the structure of posterior into the algorithm. The derivation details are presented below.

F.1 Sparse Gaussian Processes

Given data $X = \{x_i\}_{i=1}^n, \quad x_i \in \mathbb{R}^{d\times 1}$ and $y = \{y_i\}_{i=1}^n$. The sparse GP introduce a set of inducing variables, $Z = \{z_i\}_{i=1}^m, \quad z_i \in \mathbb{R}^{d\times 1}$ and the model is specified as

$$\begin{aligned} p(y_n|\mathbf{u}, Z) &= \mathcal{N}(y_n|K_{nm}K_{mm}^{-1}\mathbf{u}, \tilde{K}) \\ p(\mathbf{u}|Z) &= \mathcal{N}(\mathbf{u}|\mathbf{0}, K_{mm}). \end{aligned}$$

where $K_{mm} = [k(z_i, z_j)]_{i,j=1,\ldots,m}$, $K_{nm} = [k(x_i, z_j)]_{i=1,\ldots,n;j=1,\ldots,m}$. For different $\tilde{K}$, there are different sparse approximations for GPs. Please refer (Quiñonero-Candela and Rasmussen, 2005) for details. We test algorithms on the sparse GP model with $\tilde{K} = \beta^{-1}I$. We modify the stochastic variational inference for Gaussian processes (Hensman et al., 2013) for this model. We also apply our algorithm on the same model. However, it should be noticed that our algorithm could be easily extended to other sparse approximations (Quiñonero-Candela and Rasmussen, 2005).

We treat the inducing variables as the latent variables with uniform prior in sparse Gaussian processes. Then, the posterior of $Z, \mathbf{u}$ could be thought as the solution to the optimization problem

$$\min_{q(Z,\mathbf{u})} \int q(Z,\mathbf{u}) \log \frac{q(Z,\mathbf{u})}{p(Z)p(\mathbf{u})} \mathbf{u} dZ - \sum_{i=1}^n \int q(Z,\mathbf{u}) \log p(y_i|x_i, \mathbf{u}, Z) d\mathbf{u} dZ \tag{16}$$

The stochastic gradient of Eq.(16) w.r.t. $q(Z, \mathbf{u})$ will be

$$g(q(Z,\mathbf{u})) = \frac{1}{n}\log q(Z,\mathbf{u}) - \frac{1}{n}\log p(Z)p(\mathbf{u}) - \log p(y_i|x_i,\mathbf{u},Z)$$

and therefore, the prox-mapping in t-step is

$$\min_{q(Z,\mathbf{u})} \int q(Z,\mathbf{u}) \log \frac{q(Z,\mathbf{u})}{q_t(Z,\mathbf{u})^{1-\gamma_t/n} p(Z,\mathbf{u})^{\gamma_t/n}} \mathbf{u} dZ - \gamma_t \int q(Z,\mathbf{u}) \log p(y_i|x_i,\mathbf{u},Z) d\mathbf{u} dZ$$

which could be re-written as

$$\begin{aligned} &\min_{q(Z)q(\mathbf{u}|Z)} \int q(Z) \Bigg\{ \log \frac{q(Z)}{q_t(Z)^{1-\gamma_t/n} p(Z)^{\gamma_t/n}} \\ &+ \underbrace{\int q(\mathbf{u}|Z) \left[\log \frac{q(\mathbf{u}|Z)}{q_t(\mathbf{u}|Z)^{1-\gamma_t/n} p(\mathbf{u}|Z)^{\gamma_t/n}} - \gamma_t \log p(y_i|x_i,\mathbf{u},Z) \right] d\mathbf{u}}_{L(q(\mathbf{u}|Z))} \Bigg\} dZ \end{aligned}$$

We update $q_{t+1}(\mathbf{u}|Z)$ to be the optimal of $L(q(\mathbf{u}|Z))$ as

$$\begin{aligned} q_{t+1}(\mathbf{u}|Z) &\propto q_t(\mathbf{u}|Z)^{1-\gamma_t/n} p(\mathbf{u}|Z)^{\gamma_t/n} p(y_i|x_i,\mathbf{u},Z)^{\gamma_t} \\ &= \mathcal{N}(\mathbf{u}|m_t, \delta_t^{-1})^{1-\gamma_t/n} \mathcal{N}(\mathbf{u}|\mathbf{0}, K_{mm})^{\gamma_t/n} \mathcal{N}(y_i|K_{im}K_{mm}^{-1}\mathbf{u}, \Gamma)^{\gamma_t} \\ &= \mathcal{N}(\mathbf{u}|m_{t+1}, \delta_{t+1}^{-1}) \end{aligned}$$

where $\Gamma = diag(\tilde{K}_{ii} - Q_{ii}) + \beta^{-1}I,\ Q_{ii} = K_{im}K_{mm}^{-1}K_{mi}$,

$$\delta_{t+1} = (1-\gamma_t/n)\delta_t + \gamma_t/nK_{mm}^{-1} + \gamma_t K_{im}K_{mm}^{-1}\Gamma^{-1}K_{mm}^{-1}K_{mi}$$

$$m_{t+1} = \delta_{t+1}^{-1}\Big((1-\gamma_t/n)\delta_t^{-1}m_t + \gamma_t/nK_{mm}^{-1}m_0 + \gamma_t K_{mm}^{-1}K_{mi}\Gamma^{-1}y\Big)$$

Plug this into the $L(q(\mathbf{u}|Z))$, we have

$$L(q(u|Z)) = \int q(\mathbf{u}|Z) \left[\log \frac{q(\mathbf{u}|Z)}{q_t(\mathbf{u}|Z)^{1-\gamma_t/n} p(\mathbf{u}|Z)^{\gamma_t/n}} - \gamma_t \log p(y_i|x_i,\mathbf{u},Z) \right] d_{\mathbf{u}} = -\log \tilde{p}(y_i|x_i, Z)$$

where

$$\begin{aligned}
\tilde{p}(y_i|x_i, Z) &= \int q_t(\mathbf{u}|Z)^{1-\gamma_t/n} p(\mathbf{u}|Z)^{\gamma_t/n} p(y_i|x_i, \mathbf{u}, Z)^{\gamma_t} d\mathbf{u} \\
&= \int \mathcal{N}(\mathbf{u}|m_t, \delta_t^{-1})^{1-\gamma_t/n} \mathcal{N}(\mathbf{u}|0, K_{mm})^{\gamma_t/n} \mathcal{N}(y_i|K_{im}K_{mm}^{-1}\mathbf{u}, \Gamma)^{\gamma_t} d\mathbf{u} \\
&= \mathcal{N}(y_i|K_{im}K_{mm}^{-1}c, \Sigma)
\end{aligned}$$

where

$$\begin{aligned}
\bar{\delta}_{t+1} &= (1-\gamma_t/n)\delta_t + \gamma_t/n K_{mm}^{-1} \\
c &= \bar{\delta}_{t+1}^{-1}\Big((1-\gamma_t/n)\delta_t m_t + \gamma_t/n K_{mm}^{-1} m_0\Big) \\
\Sigma &= K_{im}K_{mm}^{-1}\bar{\delta}_{t+1}^{-1}K_{mm}^{-1}K_{mi} + \frac{1}{\gamma_t}\Gamma
\end{aligned}$$

Solve

$$\min_{q(Z)} \int q(Z) \log \frac{q(Z)}{q_t(Z)^{1-\gamma_t/n} p(Z)^{\gamma_t/n}} dZ - \int q(Z) \log \tilde{p}(y_i|x_i, Z) dZ$$

will result the update rule for $q(Z)$,

$$q_{t+1}(Z) \propto q_t(Z)^{1-\gamma_t/n} p(Z)^{\gamma_t/n} \tilde{p}(y_i|x_i, Z)$$

We approximate the $q(Z)$ with particles, i.e., $q(Z) = \sum_{j=1}^{l} w^j \delta(Z^j)$. The update rule for w^j is

$$w_{t+1}^j = \frac{w_t^j \exp(-\gamma_t/n \log(w_t^j) + \gamma_t/n \log p(Z^j) + \log \tilde{p}(y_i|x_i, Z^j))}{\sum_j^l w_t^j \exp(-\gamma_t/n \log(w_t^j) + \gamma_t/n \log p(Z^j) + \log \tilde{p}(y_i|x_i, Z^j))}$$

F.2 Latent Dirichlet Allocations

In LDA, the topics $\Phi \in \mathbb{R}^{K\times W}$ are K distributions on the words W in the text corpora. The text corpora contains D documents, the length of the d-th document is N_d. The document is modeled by a mixture of topics, with the mixing proportion $\theta_d \in \mathbb{R}^{1\times K}$. The words generating process for X_d is following: first drawing a topic assignment z_{dn}, which is 1-by-K indicator vector, *i.i.d.*from θ_d for word x_{dn} which is 1-by-W indicator vector, and then drawing the word x_{dn} from the corresponding topic $\Phi_{z_{dn}}$. We denote $z_d = \{z_{dn}\}_{n=1}^{N_d} \in \mathbb{R}^{N_d\times K}$, $x_d = \{x_{dn}\}_{n=1}^{N_d} \in \mathbb{R}^{N_d\times W}$ and $X = \{x_d\}_{d=1}^{D}$,$Z = \{Z_d\}_{d=1}^{D}$. Specifically, the joint probability is

$$\begin{aligned}
p(x_d, z_d, \theta_d, \Phi) &= p(x_d|z_d, \Phi)p(z_d|\theta_d)p(\theta_d)p(\Phi) \qquad (17) \\
p(x_d|z_d, \Phi) &= \prod_{n=1}^{N_d}\prod_{w=1}^{W}\prod_{k=1}^{K} \Phi_{kw}^{z_{dnk}x_{dnw}} \\
p(z_d|\theta_d) &= \prod_{n=1}^{N_d}\prod_{k=1}^{K} \theta_{dk}^{z_{dnk}}
\end{aligned}$$

The $p(\Phi)$ and $p(\theta)$ are the priors for parameters, $p(\theta_d|\alpha) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\prod_k^K \theta_{dk}^{\alpha-1}$ and $p(\Phi|\beta_0) = \prod_k^K \frac{\Gamma(W\beta_0)}{\Gamma(\beta_0)^W}\prod_w^W \Phi_{wk}^{\beta_0-1}$, both are Dirichlet distributions.

We incorporate the special structure into the proposed algorithm. Instead of modeling the $p(\Phi)$ solely, we model the $Z = \{Z\}_{d=1}^{D}$ and Φ together as $q(Z, \Phi)$. Based on the model, given Z, the $q(\Phi|Z)$ will be Dirichlet distribution and could be obtained in closed-form.

The posterior of Z, Φ is the solution to

$$\min_{q(Z,\Phi)} \frac{1}{D}\int q(Z,\Phi) \log \frac{q(Z,\Phi)}{p(Z|\alpha)p(\Phi|\beta)} dZd\Phi - \frac{1}{D}\sum_{d=1}^{D}\int q(Z,\Phi)\log p(x_d|z_d,\Phi)dZd\Phi$$

We approximate the finite summation by expectation, then the objective function becomes

$$\min_{q(Z,\Phi)} \frac{1}{D}\int q(Z,\Phi) \log \frac{q(Z,\Phi)}{p(Z|\alpha)p(\Phi|\beta)} dZd\Phi - \mathbb{E}_x\left[\int q(Z,\Phi)\log p(x_d|z_d,\Phi)dZd\Phi\right] \qquad (18)$$

We approximate the $q(Z) \approx \sum_{i=1}^m w^i \delta(Z^i)$ by particles, and therefore, $q(Z,\Phi) \approx \sum_{i=1}^m w^i P(\Phi|Z^i)$ where $P(\Phi|Z^i)$ is the Dirichlet distribution as we discussed. It should be noticed that from the objective function, we do not need to instantiate the z_d until we visit the x_d. By this property, we could first construct the particles $\{Z^i\}_{i=1}^m$ 'conceptually' and assign the value to $\{z_d^i\}_{i=1}^m$ when we need it. The gradient of Eq.(18) w.r.t. $q(\Phi, Z)$ is

$$g(q(Z,\Phi)) = \frac{1}{D}\log q(Z,\Phi) - \frac{1}{D}\log p(\Phi)p(Z) - \mathbb{E}_x[\log p(x_d|\Phi, z_d)]$$

Then, the SGD prox-mapping is

$$\min_{q(Z,\Phi)} \int q(Z,\Phi)\log\frac{q(Z,\Phi)}{q_t(Z,\Phi)} + \gamma_t \int q(Z,\Phi)\bigg[\log q_t(Z,\Phi)/D - \log p(\Phi)p(Z)/D - \log p(x_d|\Phi,z_d)\bigg]dZd\Phi$$

We rearrange the prox-mapping,

$$\begin{array}{rl} \min\limits_{q(Z)q(\Phi|Z)} & \displaystyle\int q(Z)q(\Phi|Z)\log\frac{q(Z)q(\Phi|Z)}{q_t(Z)^{1-\gamma_t/D}q_t(\Phi|Z)^{1-\gamma_t/D}} \\ - & \displaystyle\gamma_t\int q(Z)q(\Phi|Z)\bigg[\log p(\Phi)p(Z)/D + \log p(x_d|\Phi,z_d)\bigg]dZd\Phi \end{array}$$

$$\begin{array}{rl} \min\limits_{q(Z)q(\Phi|Z)} & \displaystyle\int q(Z)\bigg\{\log\frac{q(Z)}{q_t(Z)^{1-\gamma_t/D}p(Z)^{\gamma_t/D}} \\ + & \displaystyle\underbrace{\int q(\Phi|Z)\bigg[\log\frac{q(\Phi|Z)}{q_t(\Phi|Z)^{1-\gamma_t/D}p(\Phi)^{\gamma_t/D}} - \gamma_t\log p(x_d|\Phi,z_d)\bigg]d\Phi}_{L(q(\Phi|Z))}\bigg\}dZ \end{array}$$

The stochastic functional gradient update for $q(\Phi|Z^i)$ is

$$q_{t+1}(\Phi|Z^i) \propto q_t(\Phi|Z^i)^{1-\gamma_t/D}p(\Phi)^{\gamma_t/D}p(x_d|\Phi,z_d)^{\gamma_t}$$

Let $q_t(\Phi|Z^i) = \mathcal{D}ir(\beta_t^i)$, then, the $q_{t+1}(\Phi|Z^i)$ is also Dirichlet distribution

$$q_{t+1}(\Phi|Z^i) \propto \mathcal{D}ir(\beta_t^i)^{1-\tilde\gamma_t}\mathcal{D}ir(\beta_0)^{\tilde\gamma_t}\left(\prod_k\prod_w \Phi_{kw}^{\sum_n^{N_d}\delta(z_{dnk}=1,x_{dnw}=1)}\right)^{D\tilde\gamma_t} = \mathcal{D}ir(\beta_{t+1}^i)$$

where $\tilde\gamma_t = \gamma_t/D$ and

$$[\beta_{t+1}^i]_{kw} = (1-\tilde\gamma_t)[\beta_t^i]_{kw} + \tilde\gamma_t\beta_0 + D\tilde\gamma_t\sum_n^{N_d}\delta(z_{dnk}=1, x_{dnw}=1).$$

In mini-batch setting, the updating will be

$$[\beta_{t+1}^i]_{kw} = (1-\tilde\gamma_t)[\beta_t^i]_{kw} + \tilde\gamma_t\beta_0 + \frac{D}{B}\tilde\gamma_t\sum_{d=1}^B\sum_n^{N_d}\delta(z_{dnk}=1, x_{dnw}=1).$$

Plug the $q_{t+1}(\Phi|Z^i)$ into prox-mapping, we have

$$\begin{array}{rcl} L(q(\Phi|Z)) & = & \displaystyle\int q(\Phi|Z)\bigg[\log\frac{q(\Phi|Z)}{q_t(\Phi|Z)^{1-\tilde\gamma_t}p(\Phi)^{\tilde\gamma_t}} - D\tilde\gamma_t\log p(x_d|\Phi,z_d)\bigg]d\Phi \\ & = & -\log\tilde p(x_d|z_d,Z) \end{array}$$

where $\tilde p(x_d|z_d,Z^i) = \int_\Phi q_t(\Phi|Z^i)^{1-\tilde\gamma_t}p(\Phi)^{\tilde\gamma_t}p(x_d|\Phi,z_d)^{D\tilde\gamma_t}d\Phi$ which have closed-form

$$\begin{array}{rcl} \tilde p(x_d|z_d,Z^i) & = & \displaystyle\int_\Phi q_t(\Phi|Z^i)^{1-\tilde\gamma_t}p(\Phi)^{\tilde\gamma_t}p(x_d|\Phi,z_d)^{D\tilde\gamma_t}d\Phi \\ & = & \displaystyle\int \mathcal{D}ir(\beta_t^i)^{1-\tilde\gamma_t}\mathcal{D}ir(\beta_0^i)^{\tilde\gamma_t}\left(\prod_k\prod_w\Phi_{kw}^{\sum_n^{N_d}\delta(z_{dnk}=1,x_{dnw}=1)}\right)^{D\tilde\gamma_t}d\Phi \\ & = & \displaystyle\prod_k\left(\frac{\Gamma(\sum_w^W[\beta_t^i]_{kw})}{\prod_w\Gamma([\beta_t^i]_{kw})}\right)^{1-\tilde\gamma_t}\left(\frac{\Gamma(W\beta_0)}{\Gamma(\beta_0)^W}\right)^{\tilde\gamma_t}\frac{\prod_w\Gamma([\beta_{t+1}^i]_{kw})}{\Gamma(\sum_w[\beta_{t+1}^i]_{kw})} \end{array}$$

and

$$\begin{aligned}\log \tilde{p}(x_d|z_d, Z^i) \quad &\propto \quad \sum_k \bigg((1-\tilde{\gamma}_t) \log \Gamma(\sum_w^W [\beta_t^i]_{kw}) + \sum_w \log \Gamma([\beta_{t+1}^i]_{kw}) \\ &- \quad \log \Gamma(\sum_w [\beta_{t+1}^i]_{kw}) - (1-\tilde{\gamma}_t) \sum_w \log \Gamma([\beta_t^i]_{kw}) \bigg)\end{aligned}$$

Then, we could update $q_t(Z) = \sum_i^m w^i \delta(Z^i)$ by

$$q_{t+1}(Z^i) \propto q_t(Z^i) \exp\bigg(-\frac{\gamma_t}{D} \log q_t(Z^i) + \frac{\gamma_t}{D} \log p(Z^i|\alpha) + \log \tilde{p}(x_d|z_d, Z^i) \bigg)$$

If we set $\alpha = 1$, $p(Z^i)$ will be uniformly distributed which has no effect to the update. For general setting, to compute $\log p(Z^i|\alpha)$, we need prefix all the $\{z_d^i\}_{d=1}^D$. However, when D is huge, the second term will be small and we could ignore it approximately.

Till now, we almost complete the algorithm except the how to assign z_d when we visit x_d. We could assign the z_d randomly. However, considering the requirement for the z_d^i assignment that the $q(z_d^i|Z^i_{\setminus d}) > 0$, which means the assignment should be consistent, an better way is using the average or sampling proportional to $\int p(x_d|\Phi, z_d) q_t(\Phi|Z^i) p(z_d|Z^i_{1\ldots,d-1}) d\Phi$ where $p(z_d|Z^i_{1\ldots,d-1}) = \int p(z_d|\alpha) p(\alpha|Z^i_{1\ldots,d-1}) d\alpha$, or $\int p(x_d|\Phi, z_d) q_t(\Phi|Z^i) p(z_d|\alpha) d\Phi$.

G More Related Work

Besides the most related two inference algorithms we discussed in Section (5), *i.e.*, stochastic variational inference (Hoffman et al., 2013) and static sequential Monte Carlo (Chopin, 2002; Balakrishnan and Madigan, 2006), there are several other inference algorithms connect to the PMD from algorithm, stochastic approximation, or representation aspects, respectively.

From algorithmic aspect, our algorithm scheme shares some similarities to annealed importance sampling (AIS) (Neal, 2001) in the sense that both algorithms are sampling from a series of densities and reweighting the samples to approximate the target distribution. The most important difference is the way to construct the intermediate densities. In AIS, the density at each iteration is a weighted product of the joint distribution of *all the data* and a *fixed* proposal distribution, while the densities in PMD are a weighted product of *previous step solution* and the stochastic functional gradient on *partial data*. Moreover, the choice of the temperature parameter (fractional power) in AIS is heuristic, while in our algorithm, we have a principle way to select the stepsize with quantitative analysis. The difference in intermediate densities results the sampling step in these two algorithms is also different: the AIS might need MCMC to generate samples from the intermediate densities, while we only samples from a KDE which is more efficient. These differences make our method could handle large-scale dataset while AIS cannot.

Sequential Monte-Carlo sampler (Del Moral et al., 2006) provides a unified view of SMC in Bayesian inference by adopting different forward/backward kernels, including the variants proposed in (Chopin, 2002; Balakrishnan and Madigan, 2006) as special cases. There are subtle and important differences between the PMD and the SMC samplers. In the SMC samplers, the introduced finite forward/backward Markov kernels are used to construct a distribution over the auxiliary variables. To make the SMC samplers valid, it is required that the marginal distribution of the constructed density by integrating out the auxiliary variables must be the *exact* posterior. However, there is no such requirement in PMD. In fact, the PMD algorithm only *approaches* the posterior with controllable error by iterating the dataset *many times*. Therefore, although the proposed PMD and the SMC sampler bare some similarities operationally, they are essentially different algorithms.

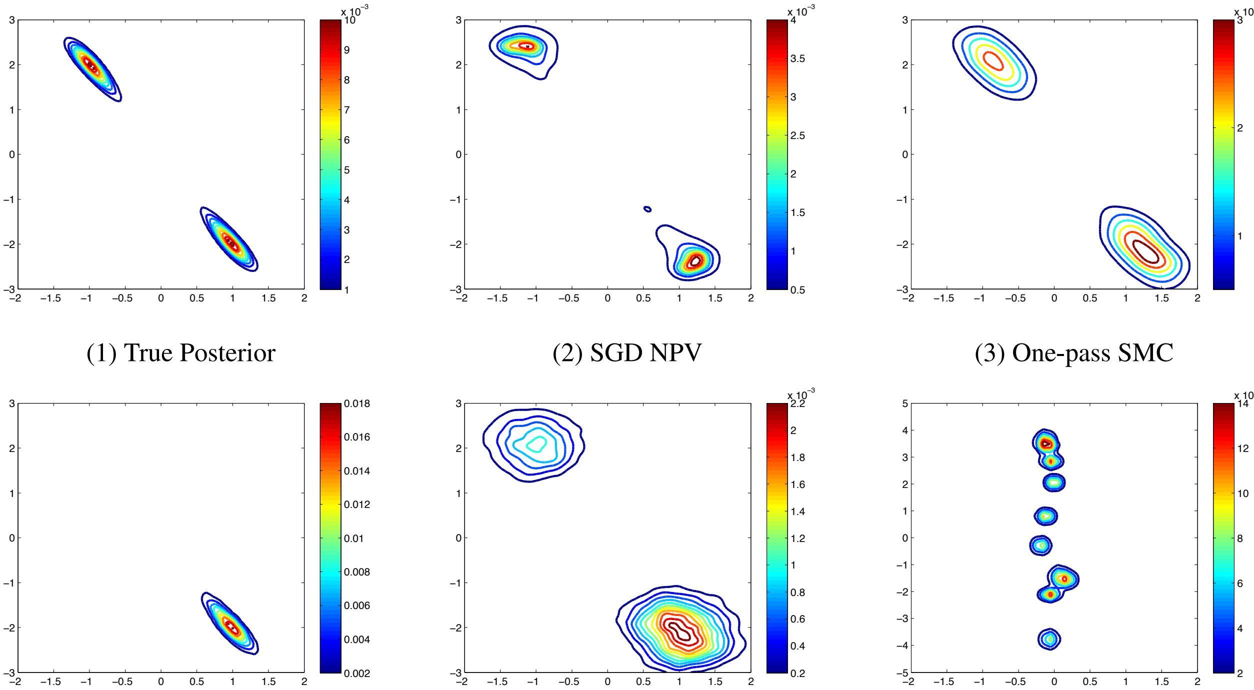
Stochastic approximation becomes a popular trick in extending the classic Bayesian inference methods to large-scale datasets recently. Besides stochastic variational inference, which incorporates stochastic gradient descent into variational inference, the stochastic gradient Langevin dynamics (SGLD) Welling and Teh (2011), and its derivatives (Ahn et al., 2012; Chen et al., 2014; Ding et al., 2014) combine ideas from stochastic optimization and Hamiltonian Monte Carlo sampling. Although both PMD and the SGLD use the stochastic gradient information to guide next step sampling, the optimization variable in these two algorithms are different which results the completely different updates and properties. In PMD, we directly update the density utilizing *functional gradient in density space*, while the SGLD perturbs the *stochastic gradient in parameter space*. Because of the difference in optimization variables, the mechanism of these algorithms are totally different. The SGLD generates a trajectory of *dependent* samples whose stationary distribution approximates the posterior, the PMD keeps an approximation of the posterior represented by *independent* particles or their weighted kernel density estimator. In fact, their different properties we discussed in Table 1 solely due to this essential difference.

A number of generalized variational inference approaches are proposed trying to relax the constraints on the density space with flexible densities. Nonparametric density family is a natural choice[2]. (Song et al., 2011) and (Ihler and McAllester, 2009; Lienart et al., 2015) extend the belief propagation algorithm with nonparametric models by kernel embedding and particle approximation, respectively. The most important difference between these algorithms and PMD is that they originate from different sources and are designed for different settings. Both the kernel BP Song et al. (2011) and particle BP Ihler and McAllester (2009); Lienart et al. (2015) are based on belief propagation optimizing *local objective* and designed for the problem with *one sample* X in which observations are highly dependent, while the PMD is optimizing the *global objective*, therefore, more similar to mean-field inference, for the inference problems with *many i.i.d.* samples.

After the comprehensive review about the similarities and differences between PMD and the existing related approximate Bayesian inference methods from algorithm, stochastic approximation and representation perspectives, we can see the position of the proposed PMD clearly. The PMD connects variation inference and Monte Carlo approximation, which seem two orthogonal paradigms in approximate Bayesian inference, and achieves a balance in trade-off between efficiency, flexibility and provability.

H Experiments Details

H.1 Mixture Models



(1) True Posterior

(2) SGD NPV

(3) One-pass SMC

(4) Gibbs Sampling

(5) SGD Langevin

(6) SVI

Figure 3: Visualization of posteriors of mixture model on synthetic dataset obtained by several inference methods.

We use the normalized Gaussian kernel in this experiment. For one-pass SMC, we use the suggested kernel bandwidth in (Balakrishnan and Madigan, 2006). For our method, since we increase the samples, the kernel bandwidth is shrunk as the theorem suggested. The batch size for stochastic algorithms and one-pass SMC is set to be 10. The total number of particles for the Monte Carlo based competitors, *i.e.*, SMC, SGD Langevin, Gibbs sampling, and our method is 1500 in total. We also keep 1500 Gaussian components in SGD NPV. The burn-in period for Gibbs sampling and stochastic Langevin dynamics are 50 and 1000 respectively.

[2] Although (Sudderth et al., 2003; Gershman et al., 2012) named their methods as "nonparametric" belief propagation and "nonparametric" variational inference, they indeed use mixture of Gaussians, which is still a parametric model.

The visualization of 10 runs average posteriors obtained by the alternative methods are plotted in Figure 3. From these figures, we could have a direct understand about the behaviors for each competitors. The Gibbs sampling and stochastic gradient Langevin dynamics sampling stuck in one local mode in each run. Gibbs sampler could fit one of the contour quite well, better than the stochastic Langevin dynamics. It should be noticed that this is the average solution, the two contours in the result of stochastic gradient Langevin dynamics did not mean it finds both modes simultaneously. The one-pass sequential Monte Carlo and stochastic nonparametric variational inference are able to location multiple modes. However, their shapes are not as good as ours. Because of the multiple modes and the highly dependent variables in posterior, the stochastic variational inference fails to converge to the correct modes.

To compare these different kinds of algorithms in a fair way, we evaluate their performances using total variation and cross entropy of the solution against the true potential functions versus the number of observations visited. In order to evaluate the total variation and the cross entropy between the true posterior and the estimated one, we use both kernel density estimation and Gaussian estimation to approximate the posterior density and report the better one for Gibbs sampling and stochastic Langevin dynamics. The kernel bandwidth is set to be 0.1 times the median of pairwise distances between data points (median trick).

In Figure 1(3)(4), the one-pass SMC performs similar to our algorithm at beginning. However, it cannot utilize the dataset effectively, therefore, it stopped with high error. It should be noticed that the one-pass SMC starts with more particles while our algorithm only requires the same number of particles at final stage. The reason that Gibbs sampling and the stochastic gradient Langevin dynamics perform worse is that they stuck in one mode. It is reasonable that Gibbs sampling fits the single mode better than stochastic gradient Langevin dynamics since it generates one new sample by scanning the whole dataset. For the stochastic nonparametric variational inference, it could locate both modes, however, it optimizes a non-convex objective which makes its variance much larger than our algorithm. The stochastic variational inference fails because of the highly dependent variables and multimodality in posterior.

H.2 Bayesian Logistic Regression

The likelihood function is

$$p(y|x,w) = \frac{1}{1+\exp(-yw^\top x)}$$

with w as the latent variables. We use Gaussian prior for w with identity covariance matrix.

We first reduce the dimension to 50 by PCA. The batch size is set to be 100 and the step size is set to be $\frac{1}{100+\sqrt{t}}$. We stop the stochastic algorithms after they pass through the whole dataset 5 times. The burn-in period for stochastic Langevin dynamic is set to be 1000. We rerun the experiments 10 times.

Although the stochastic variant of nonparametric variational inference performs comparable to our algorithm with fewer components, its speed is bottleneck when applied to large-scale problems. The gain from using stochastic gradient is dragged down by using L-BFGS to optimize the second-order approximation of the evidence lower bound.

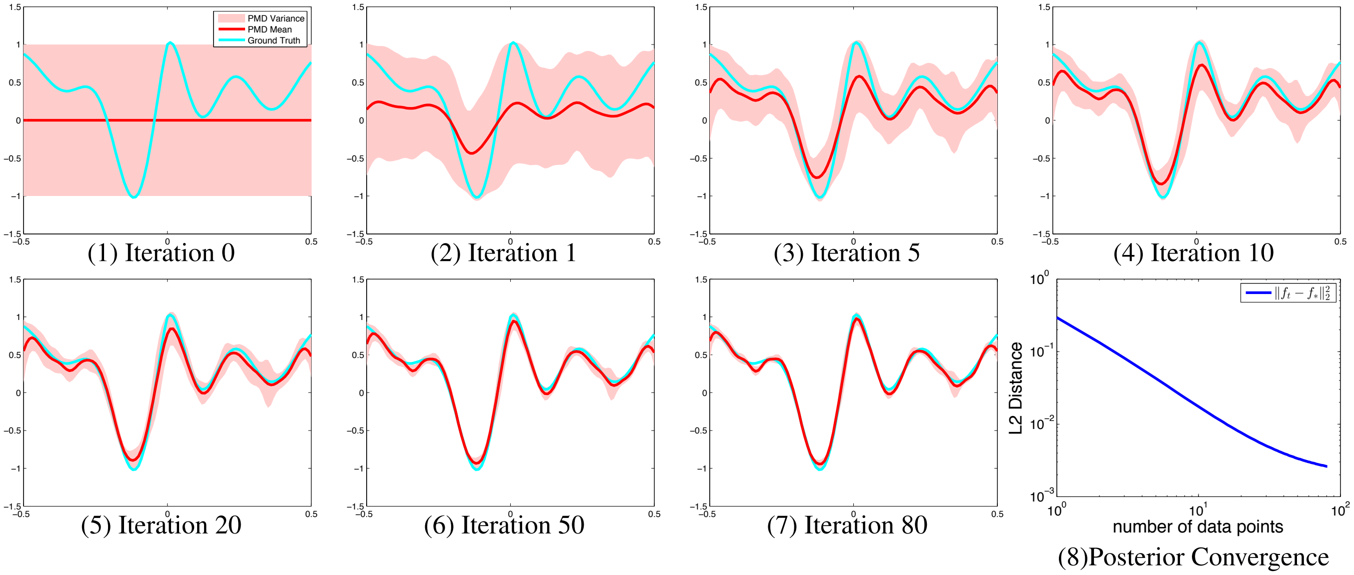
H.3 Sparse Gaussian Processes

H.3.1 1D Synthetic Dataset

We test the proposed algorithm on 1D synthetic data. The data are generated by

$$y = 3x^2 + (\sin(3.53\pi x) + \cos(7.7\pi x))\exp(-1.6\pi|x|) + 0.1e$$

where $x \in [-0.5, 0.5]$ and $e \sim \mathcal{N}(0,1)$. The dataset contains 2048 observations which is small enough to run the exact GP regression. We use Gaussian RBF kernel in Gaussian processes and sparse Gaussian processes. Since we are comparing different inference algorithms on the same model, we use the same hyperparameters for all the inference algorithms. We set the kernel bandwidth σ to be 0.1 times the median of pairwise distances between data points (median trick), and $\beta^{-1} = 0.001$. We set the stepsize in the form of $\frac{\eta}{n_0+\sqrt{t}}$ for both PMD and SVI and the batch size to be 128. Figure. 4 illustrates the evolving of the posterior provided by PMD with 16 particles and 128 inducing variables when the algorithms visit more and more data. To illustrate the convergence of the posterior provided by PMD, we initialize the $\mathbf{u} = 0$ in PMD. Later, we will see we could make the samples in PMD more efficient.



(1) Iteration 0 (2) Iteration 1 (3) Iteration 5 (4) Iteration 10

(5) Iteration 20 (6) Iteration 50 (7) Iteration 80 (8)Posterior Convergence

Figure 4: Visualization of posterior prediction distribution. The red curve is the mean function and the pale red region is the variance of the posterior. The cyan curve the ground truth. The last one shows convergence of the posterior mean to the ground truth.

H.3.2 Music Year Prediction

We randomly selected $463,715$ songs to train the model and test on $5,163$ songs. As in (Bertin-Mahieux et al., 2011), the year values are linearly mapped into $[0,1]$. The data is standardized before regression. Gaussian RBF kernel is used in the model. Since we are comparing the inference algorithms, for fairness, we fixed the model parameters for all the inference algorithms, *i.e.*, the kernel bandwidth is set to be the median of pairwise distances between data points and the observations precision $\beta^{-1} = 0.01$. We set the number of inducing inputs to be 2^{10} and batch size to be 512. The stepsize for both PMD and SVI are in the form of $\frac{\eta}{n_0+\sqrt{t}}$. To demonstrate the advantages of PMD comparing to SMC, we initialize PMD with prior while SMC with the SoD solution. We rerun the experiments 10 times. We use both 16 particles in SMC and PMD. We stop the stochastic algorithms after they pass through the whole dataset 2 times.

H.4 Latent Dirichlet Allocation

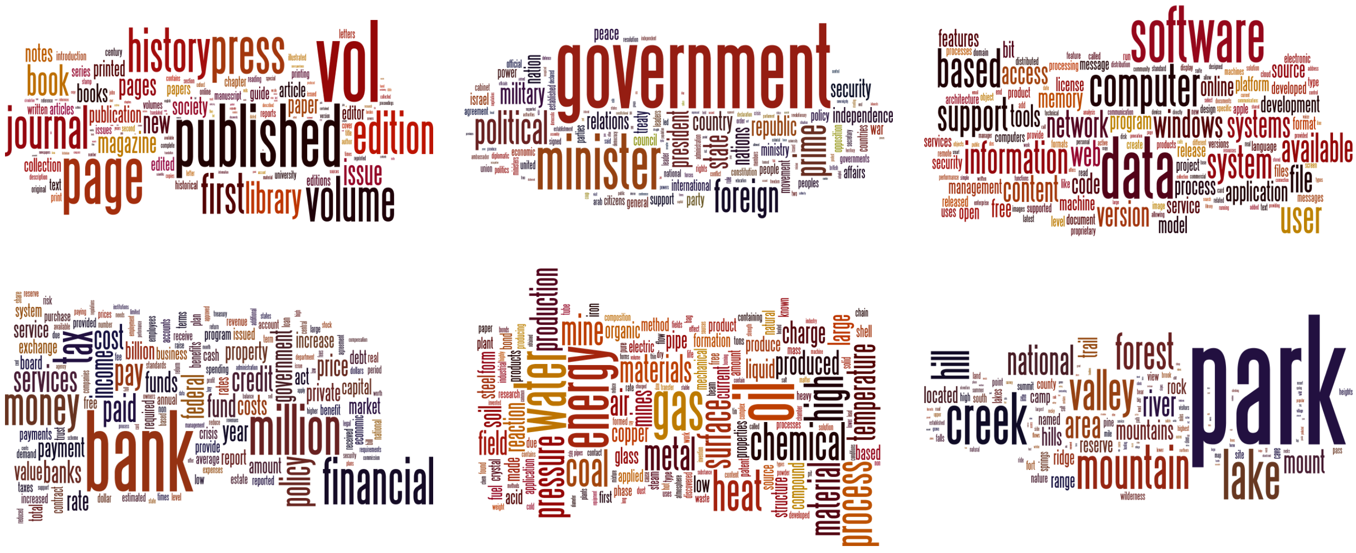


Figure 5: Several topics learnd by LDA with PMD

We fix the hyper-parameter $\alpha = 0.1$, $\beta = 0.01$, and $K = 100$. The batchsize is set to be 100. We use stepsize $\frac{\eta}{n_0 + t^\kappa}$ for PMD, stochastic variational inference and stochastic Riemannian Langevin dynamic. For each algorithm a grid-search was run on step-size parameters and the best performance is reported. We stop the stochastic algorithms after they pass through the whole dataset 5 times.

The log-perplexity was estimated using the methods discussed in (Patterson and Teh, 2013) on a separate holdout set with 1000 documents. For a document x_d in holdout set, the perplexity is computed by

$$\text{perp}(x_d|X, \alpha, \beta) = \exp\left(- \frac{\sum_{n=1}^{N_d} \log p(x_{dn}|X, \alpha, \beta)}{N_d} \right)$$

where

$$p(x_{dn}|X, \alpha, \beta) = \mathbb{E}_{\theta_d, \Phi}\left[\theta_d^\top \Phi_{\cdot, x_{dn}} \right]. \tag{19}$$

We separate the documents in testing set into two non-overlapped parts, $x_d^{\text{estimation}}$ and $x_d^{\text{evaluation}}$. We first evaluate the θ_d based on the $x_d^{\text{estimation}}$. For different inference methods, we use the corresponding strategies in learning algorithm to obtain the distribution of θ_d based on $x_d^{\text{estimation}}$. We evaluate $p(x_{dn}|X, \alpha, \beta)$ on $x_d^{\text{evaluation}}$ with the obtained distribution of θ_d. Specifically,

$$p(x_{dn}^{\text{evaluation}}|X, \alpha, \beta) = \mathbb{E}_{\Phi|X,\beta} \mathbb{E}_{\theta_d^{\text{evaluation}}|\Phi, \alpha, x_d^{\text{estimation}}} \left[\theta_d^\top \Phi_{\cdot, x_{dn}} \right]$$

For PMD, SMC and stochastic Langevin dynamics,

$$\theta_{dk}^{\text{evaluation}} = \frac{\sum_{n=1}^{N_d^{\text{estimation}}} \delta(z_{dnk}^{\text{estimation}} = 1) + \alpha}{N_d^{\text{estimation}} + K\alpha}$$

For stochastic variational inference, $q(\theta_d)$ is updated as in the learning procedure.

We illustrate several topics learned by LDA with our algorithm in Figure.5.