Primal and Dual Analysis of Entropic Fictitious Play for Finite-sum Problems

Atsushi Nitanda [1 2] **Kazusato Oko** [3 2] **Denny Wu** [4 5] **Nobuhito Takenouchi** [1] **Taiji Suzuki** [3 2]

Abstract

The entropic fictitious play (EFP) is a recently proposed algorithm that minimizes the sum of a convex functional and entropy in the space of measures — such an objective naturally arises in the optimization of a two-layer neural network in the mean-field regime. In this work, we provide a concise primal-dual analysis of EFP in the setting where the learning problem exhibits a finite-sum structure. We establish quantitative global convergence guarantees for both the continuous-time dynamics and discrete-time & finite-particle algorithm, based on properties of a *proximal Gibbs measure* introduced in Nitanda et al. (2022). Furthermore, our primal-dual framework entails a memory-efficient particle-based implementation of the EFP update, and also suggests a connection to gradient boosting methods. We illustrate the efficiency of our novel implementation in experiments including neural network optimization and image synthesis.

1. Introduction

In this work we consider the optimization of an entropy-regularized convex functional in the space of measures:

$$\min_{\mu \in \mathcal{P}_2} \left\{ F(\mu) + \lambda \mathrm{Ent}(\mu) \right\}. \tag{1}$$

Note that when F is a linear functional, i.e., $F(\mu) = \int f \mathrm{d}\mu$, then the above objective admits a unique minimizer $\mu_* \propto \exp(-\lambda^{-1} f)$, samples from which can be obtained using various methods such as Langevin Monte Carlo. In the more general setting of nonlinear F, efficiently optimizing (1) becomes more challenging, and existing algorithms typically take the form of an *interacting-particle* update.

[1]Kyushu Institute of Technology [2]Center for Advanced Intelligence Project [3]University of Tokyo [4]University of Toronto [5]Vector Institute for Artificial intelligence. Correspondence to: Atsushi Nitanda <nitanda@ai.kyutech.ac.jp>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Solving the objective (1) is closely related to the optimization of neural networks in the *mean-field regime*, which attracts attention because it captures the representation learning capacity of neural networks (Ghorbani et al., 2019; Li et al., 2020; Abbe et al., 2022; Ba et al., 2022). A key ingredient of the mean-field analysis is the connection between gradient descent on the wide neural network and the Wasserstein gradient flow in the space of measures, based on which the global convergence of training dynamics can be shown by exploiting the convexity of the loss function (Nitanda & Suzuki, 2017; Mei et al., 2018; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2018; Sirignano & Spiliopoulos, 2020). However, most of these existing analyses do not prove a convergence rate for the studied dynamics.

The entropy term in (1) leads to a *noisy* gradient descent update, where the injected Gaussian noise encourages exploration (Mei et al., 2018; Hu et al., 2019). Recent works have shown that this added regularization entails exponential convergence of the continuous dynamics (termed the mean-field Langevin dynamics) under a *logarithmic Sobolev inequality* condition that is easily verified in regularized empirical risk minimization problems with neural networks (Nitanda et al., 2022; Chizat, 2022; Chen et al., 2022; Suzuki et al., 2023). Moreover, novel update rules that optimize (1) have also been proposed in Nitanda et al. (2021); Oko et al. (2022); Nishikawa et al. (2022), for which quantitative convergence guarantees can be shown by adapting classical convex optimization theory into the space of measures.

An important observation behind the design of these new algorithms is a *self-consistent* condition of the global optimum (Hu et al., 2019), namely, the optimal solution to the objective (1) must satisfy

$$\mu = \hat{\mu}, \quad \text{where } \hat{\mu}(x) \propto \exp\left(-\frac{1}{\lambda} \frac{\delta F(\mu)}{\delta \mu}(x) \right), \tag{2}$$

in which $\frac{\delta F}{\delta \mu}$ denotes the first-variation of F. Following Nitanda et al. (2022), we refer to $\hat{\mu}$ as the *proximal Gibbs distribution*. Based on properties of $\hat{\mu}$, Oko et al. (2022) and Nitanda et al. (2022) established a primal-dual formulation of the optimization problem in the finite-sum setting.

In this work, we analyze a different update rule that optimizes (1) recently proposed in Ren & Wang (2022) termed the *entropic fictitious play* (EFP). The EFP update is inspired

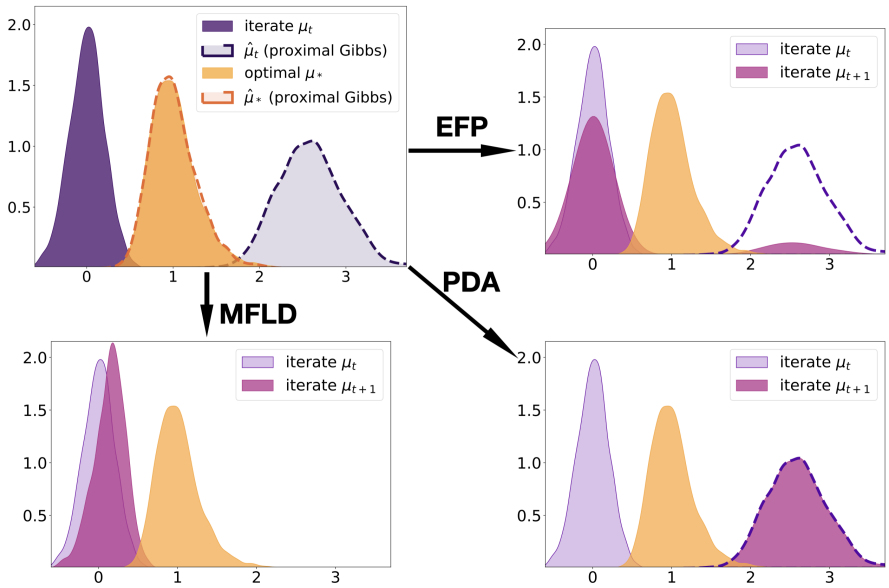

Figure 1. 1D visualization of the parameter distribution of mean-field two-layer neural network (tanh) optimized by the mean-field Langevin dynamics (MFLD), particle dual averaging[2](PDA), and entropic fictitious play (EFP), all of which can solve the self-consistent equation $\mu = \hat{\mu}$, where $\hat{\mu}$ is the *proximal Gibbs measure* associated with μ. We set $\lambda = \lambda' = 10^{-2}$.

by the classical fictitious play in game theory for learning Nash equilibria (Brown, 1951), which has been revisited by (Cardaliaguet & Hadikhanloo, 2017; Hadikhanloo & Silva, 2019; Perrin et al., 2020; Lavigne & Pfeiffer, 2022) in the context of mean-field games. Intuitively, the EFP method successively accumulates $\hat{\mu} - \mu$ to the current iteration μ with an appropriate weight, as illustrated in Figure 1. In light of the self-consistent condition (2), this iteration is quite intuitive and natural, as it can be interpreted as a fixed-point iteration on the optimality condition $\mu = \hat{\mu}$.

Despite its simple structure, EFP can be computationally prohibitive if implemented in a naive way. Specifically, for approximately optimizing mean-field models, we typically utilize a finite-particle discretization of the probability distribution. Then, the additive structure of EFP update requires additional particles to be added at each iteration. Therefore, the number of particles linearly increases in proportion to the number of steps, and nontrivial heuristics are needed to reduce the computational costs in general settings.

1.1. Our Contributions

In this paper, we present a concise primal-dual analysis of EFP in the finite-sum setting (e.g., empirical risk minimization). We establish convergence in both continuous- and discrete-time of the primal and dual objectives, which is equivalent to the convergence of the Kullback-Leibler (KL) divergence between μ and $\hat{\mu}$. Furthermore, our framework naturally suggests an efficient implementation of EFP for finite-sum problems, where the memory consumption remains constant across the iterations. Our contributions can be summarized as follows.

- We provide a simple analysis of EFP for finite-sum problems through the primal-dual framework. Our global convergence result holds in both continuous-and discrete-time settings (Theorem 4.4 and 5.4).
- Our theory suggests a memory-efficient implementation of EFP (Algorithm 2). Specifically, we do not memorize all particles from the previous iterations, but instead store the information via the proximal Gibbs distribution, with memory cost that only depends on the number of training samples but not the number of iterations.
- We present a connection between EFP and gradient-boosting methods, which enables us to establish discrete-time global convergence based on the Frank-Wolfe argument (Theorem 6.1). This analysis also allows us to characterize the discretization error due to finite particles.
- We employ our novel implementation of EFP in various experiments including neural network learning and image synthesis (Tian & Ha, 2022).

1.2. Notations

$\|\cdot\|_2$, $\|\cdot\|_\infty$ denote the Euclidean norm and uniform norm, respectively. Given a probability distribution μ on $\mathbb{R}^d$, we write the expectation w.r.t. $\theta \sim \mu$ as $\mathbb{E}_{\theta\sim\mu}[\cdot]$ or simply $\mathbb{E}_\mu[\cdot]$, $\mathbb{E}[\cdot]$ when the random variable and distri-

[2]For illustrative purposes, we simplify the PDA implementation from Nitanda et al. (2021) by removing the weighted averaging in the gradient.

bution are obvious from the context; e.g. for a function $f : \mathbb{R}^d \to \mathbb{R}$, we write $\mathbb{E}_\mu[f] = \int f(\theta)\mathrm{d}\mu(\theta)$ when f is integrable. KL stands for the Kullback-Leibler divergence: $\mathrm{KL}(\mu\|\mu') \overset{\mathrm{def}}{=} \int \mathrm{d}\mu(\theta) \log \frac{\mathrm{d}\mu}{\mathrm{d}\mu'}(\theta)$ and Ent stands for the negative entropy: $\mathrm{Ent}(\mu) = \int \mathrm{d}\mu(\theta) \log \frac{\mathrm{d}\mu}{\mathrm{d}\theta}(\theta)$. Let $\mathcal{P}_2$ be the space of probability distributions on $\mathbb{R}^d$ such that absolutely continuous with respect to Lebesgue measure and the entropy and second moment are well-defined.

2. Preliminaries

2.1. Problem Setup

We say the functional $G : \mathcal{P}_2 \to \mathbb{R}$ is differentiable and convex when G satisfies the following. First, there exists a functional (referred to as a first variation): $\frac{\delta G}{\delta \mu} : \mathcal{P}_2 \times \mathbb{R}^d \ni (\mu, \theta) \mapsto \frac{\delta G}{\delta \mu}(\mu)(\theta) \in \mathbb{R}$ such that for any $\mu, \mu' \in \mathcal{P}_2$,

$$\left.\frac{\mathrm{d}G(\mu + \epsilon(\mu' - \mu))}{\mathrm{d}\epsilon}\right|_{\epsilon=0} = \int \frac{\delta G}{\delta \mu}(\mu)(\theta)\mathrm{d}(\mu' - \mu)(\theta)$$

and G satisfies the convexity condition:

$$G(\mu') \geq G(\mu) + \int \frac{\delta G}{\delta \mu}(\mu)(\theta)\mathrm{d}(\mu' - \mu)(\theta). \tag{3}$$

Let $F_0 : \mathcal{P}_2 \to \mathbb{R}$ be differentiable and convex, and define $F(\mu) = F_0(\mu) + \lambda' \mathbb{E}_\mu[\|\theta\|_2^2]$. We consider the minimization of an entropy-regularized convex functional:

$$\min_{\mu \in \mathcal{P}_2} \left\{ \mathcal{L}(\mu) = F(\mu) + \lambda \mathrm{Ent}(\mu) \right\}. \tag{4}$$

Note both F and $\mathcal{L}$ are also differentiable convex functionals. Throughout the paper, we make the following regularity assumption on F_0.

Assumption 2.1. We assume the first variation $\frac{\delta F_0}{\delta \mu}$ satisfies the Lipschitz continuity and boundedness in the following sense: there exists constant $M > 0$ such that for any $\mu, \mu' \in \mathcal{P}_2$ and for any $\theta, \theta' \in \mathbb{R}^d$,

$$\left| \frac{\delta F_0}{\delta \mu}(\mu)(\theta) - \frac{\delta F_0}{\delta \mu}(\mu)(\theta') \right| \leq M(W_2(\mu, \mu') + \|\theta - \theta'\|_2),$$
$$\left| \frac{\delta F_0}{\delta \mu}(\mu)(\theta) \right| \leq M,$$

where W_2 is the 2-Wasserstein distance.

For $\mu \in \mathcal{P}_2$, we introduce the *proximal Gibbs distribution*, which plays a key role in our analysis (for basic properties see Nitanda et al. (2022); Chizat (2022)).

Definition 2.2 (Proximal Gibbs distribution)**.** We define $\hat{\mu}$ be the Gibbs distribution so that

$$\frac{\mathrm{d}\hat{\mu}}{\mathrm{d}\theta}(\theta) = \frac{\exp\left(-\frac{1}{\lambda}\frac{\delta F(\mu)}{\delta \mu}(\theta)\right)}{Z(\mu)}, \tag{5}$$

where $Z(\mu)$ is the normalization constant and $\mathrm{d}\hat{\mu}/\mathrm{d}\theta$ represents the density function w.r.t. Lebesgue measure.

Remark 2.3. Under Assumption 2.1, the existence and uniqueness of the minimizer $\mu_* \in \mathcal{P}_2$ of $\mathcal{L}$ is guaranteed by Ren & Wang (2022), and μ_* satisfies the first-order optimality condition: $\mu_* = \hat{\mu}_*$ (Hu et al., 2019).

We focus on the finite-sum setting as follows. Given differentiable convex functions: $\ell_i : \mathbb{R} \to \mathbb{R}$ and functions $h_i : \mathbb{R}^d \to \mathbb{R}$ ($i \in \{1, \ldots, n\}$), we consider the following minimization problem:

$$\min_{\mu \in \mathcal{P}_2} \left\{ \mathcal{L}(\mu) = \frac{1}{n}\sum_{i=1}^n \ell_i(\mathbb{E}[h_i]) + \lambda' \mathbb{E}[\|\theta\|_2^2] + \lambda \mathrm{Ent}(\mu) \right\}, \tag{6}$$

where the expectation is taken with respect to $\theta \sim \mu$, i.e., $\mathbb{E}[h_i] = \mathbb{E}_{\theta\sim\mu}[h_i(\theta)]$. Note that the problem (6) is a special case of (4) by setting $F_0(\mu) = \frac{1}{n}\sum_{i=1}^n \ell_i(\mathbb{E}_\mu[h_i(\theta)])$. In this setting, the first variation of F is

$$\frac{\delta F}{\delta \mu}(\mu)(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_i'(\mathbb{E}[h_i])h_i(\theta) + \lambda' \|\theta\|_2^2.$$

2.2. Example Applications

One noticeable example of problem (6) is the learning of mean-field neural networks via empirical risk minimization.

Example 2.4 (Mean-field model). Let $h(\theta, \cdot) : \mathcal{X} \to \mathbb{R}$ be a function parameterized by $\theta \in \mathbb{R}^d$, where $\mathcal{X}$ is the data space. The mean-field model is an integration of $h(\theta, \cdot)$ with respect to the distribution $\mu \in \mathcal{P}_2$ over the parameter space: $\mathbb{E}_{\theta\sim\mu}[h(\theta, \cdot)]$. Given training examples $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathbb{R}$, denote $h_i(\theta)$ as the output $h(\theta, x_i)$. Then we may choose the loss term as $\ell_i(\mathbb{E}_\mu[h_i(\theta)]) = \ell(y_i, \mathbb{E}[h(\theta, x_i)])$, for convex loss functions such as the squared loss $\ell(y, y') = 0.5(y - y')^2$ and the logistic loss $\log(1 + \exp(-yy'))$.

Another example that falls into the framework of (6) is density estimation using a mixture model.

Example 2.5 (Density Estimation). Let $\mu * g_\delta$ ($\mu \in \mathcal{P}_2$) be the Gaussian convolution of μ:

$$\mu * g_\delta(\cdot) \propto \int \exp\left(-\frac{\|\theta - \cdot\|_2^2}{2\sigma^2}\right) \mathrm{d}\mu(\theta),$$

where $g_\delta(\theta) = (2\pi\sigma^2)^{-d/2}\exp(-\|\theta^2\|/(2\sigma^2))$. Let $\{\zeta_i\}_{i=1}^n \subset \mathbb{R}^d$ be the i.i.d. observations. Then, the log-likelihood $-\sum_{i=1}^n \log(\mu * g_\delta(\zeta_i))$ can be written as $\sum_{i=1}^n \ell_i(\mathbb{E}_\mu[h_i])$, by setting $\ell_i = -\log$ and $h_i(\theta) = g_\delta(\theta - \zeta_i)$ since $\mu * g_\delta(\zeta_i) = \mathbb{E}_{\theta\sim\mu}[g_\delta(\theta - \zeta_i)]$.

3. Entropic Fictitious Play

3.1. The Ideal Update

The entropic fictitious play (EFP) algorithm (Ren & Wang, 2022) is the optimization method that minimizes the objec-

tive (4). The continuous evolution $\{\mu_t\}_{t\geq 0} \subset \mathcal{P}_2$ of the EFP is defined as follows: for $\gamma > 0$,

$$\frac{\mathrm{d}\mu_t}{\mathrm{d}t} = \gamma(\hat{\mu}_t - \mu_t). \tag{7}$$

Under Assumption 2.1, the existence of differentiable density functions that solve (7) is guaranteed.

The time-discretized version of the EFP can be obtained by the explicit Euler's method. Given a step-size $\eta > 0$,

$$\mu_{t+\eta} = (1-\eta\gamma)\mu_t + \eta\gamma\hat{\mu}_t. \tag{8}$$

To execute this update, we need to compute the proximal Gibbs distribution $\hat{\mu}_t$, which can be accessed via standard sampling algorithms such as Langevin Monte Carlo. This ideal update is summarized in Algorithm 1 below.

Algorithm 1 Discrete-time Entropic Fictitious Play

Input: $\mu^{(0)}$, T, η, γ
for $t = 0$ **to** $T-1$ **do**
  Compute $\hat{\mu}^{(t)}$ via standard sampling algorithms
  Update $\mu^{(t+1)} = (1-\eta\gamma)\mu^{(t)} + \eta\gamma\hat{\mu}^{(t)}$
end for
Return $\mu^{(T)}$

As previously remarked, in light of the first-order optimality condition $\mu_* = \hat{\mu}_*$, the discrete-time EFP (8) is an intuitive update that mirrors the fixed-point iteration methods. However, the naive particle implementation of (8) would be extremely expensive in terms of memory cost. In particular, the additive structure of the discrete-time EFP update requires us to store all the past particles along the trajectory. In other words, the memory cost scales linearly with the number of iterations. This prohibits the practical use of EFP beyond very short time horizon.

3.2. Efficient Particle Update for Finite-sum Settings

We now present an efficient implementation of discrete-time EFP by exploiting the finite-sum structure as in Nitanda et al. (2021); Oko et al. (2022). For comparison, we first provide a recap for the naive implementation. To execute Algorithm 1, we run a sampling algorithm for each Gibbs distribution $\hat{\mu}^{(t)}$, and build an approximation $\hat{\mu}^{(t)} \sim \frac{1}{m}\sum_{r=1}^m \delta_{\theta_r^{(t)}}$ by finite particles $\theta_r^{(t)} \sim \hat{\mu}^{(t)}$ $(r = 1, 2, \ldots, m)$, where δ_θ is the Dirac measure at $\theta \in \mathbb{R}^d$. Then, the approximation $\underline{\mu}^{(t+1)}$ to $\mu^{(t+1)}$ can be recursively constructed as follows:

$$\underline{\mu}^{(t+1)} = (1-\eta\gamma)\underline{\mu}^{(t)} + \frac{\eta\gamma}{m}\sum_{r=1}^m \delta_{\theta_r^{(t)}}. \tag{9}$$

It is clear that $\underline{\mu}^{(t)}$ consists of all particles obtained up to the t-th iteration, and thus the number of particles accumulates linearly with respect to the number of iterations.

Alternative implementation. Our starting observation is that sampling algorithms for the Gibbs distribution such as Langevin Monte Carlo only require the computation of the gradient of the potential function. Recall that the potential function $\nabla\frac{\delta F(\mu^{(t)})}{\delta\mu}(\theta)$ of the proximal Gibbs distribution $\hat{\mu}^{(t)}$ for finite-sum problems (6) is written as follows:

$$\frac{1}{n}\sum_{i=1}^n \ell_i'(\mathbb{E}_{\mu^{(t)}}[h_i(\theta)])\nabla h_i(\theta) + 2\lambda'\theta. \tag{10}$$

Hence to approximate (10), we do not need to store particles constituting $\underline{\mu}^{(t)}$ themselves, but only the empirical averages: $\mathbb{E}_{\underline{\mu}^{(t)}}[h_i(\theta)]$ $(i = 1, 2, \ldots, n)$. Then, we notice that these empirical averages can be recursively computed in an online manner, because of the additive structure of the EFP update (9): for $i = 1, 2, \ldots, n$,

$$\mathbb{E}_{\underline{\mu}^{(t+1)}}[h_i(\theta)] = (1-\eta\gamma)\mathbb{E}_{\underline{\mu}^{(t)}}[h_i(\theta)] + \frac{\eta\gamma}{m}\sum_{r=1}^m h_i(\theta_r^{(t)}).$$

It is clear that the number of particles required to compute the above update is m (i.e., the number of samples drawn from the current Gibbs distribution), which can be kept constant during training. This gives a memory-efficient and implementable EFP update, which is described in Algorithm 2 using Langevin Monte Carlo (the sampling algorithm can be replaced by other alternatives). For simplicity, we denote $H_i^{(t)} = \mathbb{E}_{\underline{\mu}^{(t)}}[h_i(\theta)]$ in the algorithm description.

Algorithm 2 Efficient Implementation of EFP for Finite-sum Problem

Input: $\mu^{(0)}$, λ, λ', T, S, m, η, γ, η'
Randomly sample m particles $\{\theta_r^{(-1)}\}_{r=1}^m$ from $\mu^{(0)}$
Compute initial empirical averages: for $i \in \{1, 2, \ldots, n\}$
$H_i^{(0)} \leftarrow \frac{1}{m}\sum_{r=1}^m h_i(\theta_r^{(-1)})$
for $t = 0$ **to** $T-1$ **do**
  Initialize $\{\overline{\theta}_r^{(0)}\}_{r=1}^m$ (e.g., $\{\overline{\theta}_r\}_{r=1}^m \leftarrow \{\theta_r^{(t-1)}\}_{r=1}^m$)
  for $s = 0$ **to** $S-1$ **do**
    Run Langevin Monte-Carlo: for $r \in \{1, 2, \ldots, m\}$
    $g^{(s)}(\overline{\theta}_r^{(s)}) \leftarrow \frac{1}{n}\sum_{i=1}^n \ell_i'(H_i^{(t)})\nabla h_i(\overline{\theta}_r^{(s)})$
    $\xi_r^{(s)} \sim \mathcal{N}(0, I_d)$ (i.i.d. Gaussian noise)
    $\overline{\theta}_r^{(s+1)} \leftarrow (1-2\eta'\lambda')\overline{\theta}_r^{(s)} - \eta' g^{(s)}(\overline{\theta}_r^{(s)}) + \sqrt{2\eta'\lambda}\xi_r^{(s)}$
  end for
  $\{\theta_r^{(t)}\}_{r=1}^m \leftarrow \{\overline{\theta}_r^{(S)}\}_{r=1}^m$
  Update empirical averages: for $i \in \{1, 2, \ldots, n\}$
  $H_i^{(t+1)} \leftarrow (1-\eta\gamma)H_i^{(t)} + \frac{\eta\gamma}{m}\sum_{r=1}^m h_i(\theta_r^{(t)})$
end for
Return $\{\theta_r^{(T-1)}\}_{r=1}^m$ and $\{H_i^{(T)}\}_{i=1}^n$

Remark 3.1. The iteration complexity of Langevin Monte Carlo has been extensively studied. In our problem setting,

we may verify that the Gibbs measure $\hat{\mu}_t$ satisfies a logarithmic Sobolev inequality (LSI) via the Holley-Stroock perturbation argument (Holley & Stroock, 1987). Therefore, convergence of LMC is guaranteed from prior results (e.g., Vempala & Wibisono (2019)).

Benefits of the proximal Gibbs distribution. When making the prediction by mean-field models (Example 2.4) in machine learning applications, we need to compute $\mathbb{E}_\mu[h(\theta, x)]$ on unseen input $x \in \mathcal{X}$ with an optimized distribution μ. In the following section, we establish primal and dual convergence of EFP, which implies the convergence of the proximal Gibbs distribution $\hat{\mu}_t$ to the optimum. Therefore, we can utilize particles $\{\theta_r^{(T-1)}\}_{r=1}^m$ to construct an actual predictor $\frac{1}{m}\sum_{r=1}^m h(\theta_r^{(T-1)}, x)$ because these particles follow the proximal Gibbs distribution which is approximately optimal. Moreover, if we want to reduce the number of particles, we can apply efficient compression methods for the Gibbs distribution such as the kernel quadrature (Bach, 2017) and herding (Chen et al., 2018) to further improve the particle complexity.

4. Convergence in Continuous Time

In this section, we provide a concise convergence analysis of the EFP in continuous-time (7) using properties of the proximal Gibbs distribution (5).

4.1. Primal Convergence

Let $\mu_* \in \mathcal{P}_2$ be an optimal solution of (4). The following property relates the suboptimality gap to the KL divergence involving the proximal Gibbs measure.

Proposition 4.1 (Nitanda et al. (2022); Chizat (2022)). *For any $\mu \in \mathcal{P}_2$, we get*

(1) $$\frac{\delta \mathcal{L}}{\delta \mu}(\mu) = \lambda \frac{\delta}{\delta \mu'} \mathrm{KL}(\mu' \| \hat{\mu})|_{\mu'=\mu} = \lambda \log \frac{\mathrm{d}\mu}{\mathrm{d}\hat{\mu}},$$

(2) $$\lambda \mathrm{KL}(\mu \| \hat{\mu}) \geq \mathcal{L}(\mu) - \mathcal{L}(\mu_*) \geq \lambda \mathrm{KL}(\mu \| \mu_*).$$

We now show that the continuous-time EFP (7) converges linearly to the global optimal solution μ_*.

Theorem 4.2. *Let $\{\mu_t\}_{t\geq 0}$ be the evolution described by (7). Under Assumption 2.1, we get for $t \geq 0$,*

$$\mathcal{L}(\mu_t) - \mathcal{L}(\mu_*) \leq \exp(-\gamma t)(\mathcal{L}(\mu_0) - \mathcal{L}(\mu_*)).$$

Proof. From Proposition 4.1, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{L}(\mu_t) - \mathcal{L}(\mu_*)) \\
&= \int \frac{\delta \mathcal{L}}{\delta \mu}(\mu_t)(\theta) \frac{\mathrm{d}\mu_t}{\mathrm{d}t}(\theta) \\
&= \int \lambda\gamma \log \frac{\mathrm{d}\mu_t}{\mathrm{d}\hat{\mu}_t}(\theta) \mathrm{d}(\hat{\mu}_t - \mu_t)(\theta) \\
&= -\lambda\gamma \left(\mathrm{KL}(\mu_t \| \hat{\mu}_t) + \mathrm{KL}(\hat{\mu}_t \| \mu_t)\right) \\
&\leq -\gamma(\mathcal{L}(\mu_t) - \mathcal{L}(\mu_*)).
\end{aligned}$$

The statement then follows from a straightforward application of Grönwall's inequality. □

4.2. Primal-Dual Convergence

To introduce a primal-dual formulation of the finite-sum minimization problems (6), we denote $\ell_i^*(\cdot)$ as the Fenchel conjugate, i.e.,

$$\ell_i^*(z^*) = \sup_{z \in \mathbb{R}} \{z z^* - \ell_i(z)\}. \text{ for } z^* \in \mathbb{R}$$

Also, for any given vector $g = \{g_i\}_{i=1}^n \in \mathbb{R}^n$, we define

$$q_g(\theta) = \exp\left(-\frac{1}{\lambda}\left(\frac{1}{n}\sum_{i=1}^n h_i(\theta) g_i + \lambda' \|\theta\|_2^2\right)\right).$$

Then, the dual problem of (6) is defined as

$$\max_{g \in \mathbb{R}^n} \left\{ \mathcal{D}(g) = -\frac{1}{n}\sum_{i=1}^n \ell_i^*(g_i) - \lambda \log \int q_g(\theta) \mathrm{d}\theta \right\}. \tag{11}$$

The duality theorem (Rockafellar, 1970; Bauschke et al., 2011; Oko et al., 2022) guarantees the relationship $\mathcal{D}(g) \leq \mathcal{L}(\mu)$ for any $g \in \mathbb{R}^n$ and $\mu \in \mathcal{P}_2$, and it is known that the duality gap $\mathcal{L}(\mu) - \mathcal{D}(g)$ vanishes at the solutions of (6) and (11) when they exist. In our problem setting, a more precise result is given in Nitanda et al. (2022). We write $g_\mu = \{\ell_i'(\mathbb{E}_\mu[h_i(\theta)])\}_{i=1}^n \in \mathbb{R}^n$ ($\mu \in \mathcal{P}_2$). Note that q_g and g_μ are connected to the proximal Gibbs distribution through the relation: $q_{g_\mu} \propto \hat{\mu}$. The following theorem from Nitanda et al. (2022) exactly characterizes the duality gap $\mathcal{L}(\mu) - \mathcal{D}(g_\mu)$ between $\mu \in \mathcal{P}_2$ and $g_\mu \in \mathbb{R}^n$ via the KL divergence to the proximal Gibbs distribution $\hat{\mu}$.

Theorem 4.3 (Duality Theorem (Nitanda et al., 2022)). *Suppose $\ell_i(\cdot)$ is convex and differentiable. For any $\mu \in \mathcal{P}_2$, the duality gap between μ and g_μ satisfies*

$$0 \leq \mathcal{L}(\mu) - \mathcal{D}(g_\mu) = \lambda \mathrm{KL}(\mu \| \hat{\mu}).$$

For the rest of the analysis, we further assume ℓ_i is the twice differentiable convex function. The following theorem shows the exponential convergence of $\mathrm{KL}(\mu_t \| \hat{\mu}_t)$ which upper bounds the duality gap.

Theorem 4.4. *Let $\{\mu_t\}_{t\geq 0}$ be the evolution described by (7). Under Assumptions 2.1, we get*

$$\mathrm{KL}(\mu_t\|\hat{\mu}_t) \leq \exp(-\gamma t)\mathrm{KL}(\mu_0\|\hat{\mu}_0).$$

Proof. We calculate the time derivative of the primal and dual objectives. On one hand, following the same computation as the proof of Theorem 4.2,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\mu_t) \leq -\lambda\gamma\mathrm{KL}(\mu_t\|\hat{\mu}_t).$$

On the other hand, we have

$$\begin{aligned}
&-\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{D}(g_{\mu_t}) = -\nabla\mathcal{D}(g_{\mu_t})^\top\frac{\mathrm{d}g_{\mu_t}}{\mathrm{d}t}\\
&= \frac{1}{n}\sum_{i=1}^n\left(\ell_i^{*\prime}(g_{\mu_t,i}) - \frac{1}{\int q_{g_{\mu_t}}(\theta)\mathrm{d}\theta}\int h_i(\theta)q_{g_{\mu_t}}(\theta)\mathrm{d}\theta\right)\\
&\quad\cdot\frac{\mathrm{d}}{\mathrm{d}t}\ell_i'(\mathbb{E}_{\mu_t}[h_i(\theta)])\\
&= -\frac{\gamma}{n}\sum_{i=1}^n\left(\mathbb{E}_{\mu_t}[h_i(\theta)] - \mathbb{E}_{\hat{\mu}_t}[h_i(\theta)]\right)^2\ell_i''(\mathbb{E}_{\mu_t}[h_i]) \leq 0,
\end{aligned}$$

where we used $\hat{\mu}_t \propto q_{g_{\mu_t}}(\theta)\mathrm{d}\theta$, $\ell_i^{*\prime} = (\ell_i')^{-1}$, and

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\ell_i'(\mathbb{E}_{\mu_t}[h_i(\theta)]) &= \ell_i''(\mathbb{E}_{\mu_t}[h_i])\int h_i(\theta)\frac{\mathrm{d}\mu_t}{\mathrm{d}t}(\theta)\\
&= \ell_i''(\mathbb{E}_{\mu_t}[h_i])(\mathbb{E}_{\hat{\mu}_t}[h_i] - \mathbb{E}_{\mu_t}[h_i]).
\end{aligned}$$

Combining these inequalities with Theorem 4.4,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\mu_t\|\hat{\mu}_t) &= \frac{1}{\lambda}\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{L}(\mu_t) - \mathcal{D}(g_{\mu_t}))\\
&\leq -\gamma\mathrm{KL}(\mu_t\|\hat{\mu}_t).
\end{aligned}$$

Therefore, Grönwall's inequality finishes the proof. □

Finally, Theorem 4.4 in combination with the following result entails the convergence of $\mathcal{L}(\hat{\mu}_t)$ to the optimal value.

Theorem 4.5 (Nitanda et al. (2022)). *Suppose Assumptions in Theorem 4.3 holds and suppose that ℓ_i is L-Lipschitz smooth, that is, $|\ell'(z)-\ell(z')| \leq L|z-z'|$ for any $z, z' \in \mathbb{R}$, and $\|h_i\|_\infty \leq B$. Then,*

$$\mathcal{L}(\hat{\mu}_t) - \mathcal{D}(g_{\mu_t}) \leq (\lambda + 2B^2L)\mathrm{KL}(\mu_t\|\hat{\mu}_t).$$

5. Convergence in Discrete Time

Now we establish the convergence guarantee for the discrete-time EFP (Algorithm 1). We present two different proof strategies that might be of independent interest. The first approach uses a one-step interpolation argument that builds upon the continuous-time result in Section 4; whereas the second approach is a Frank-Wolfe analysis based on the gradient-boosting interpretation of EFP.

5.1. Analysis via One-step Interpolation

The key to our first convergence proof is the one-step interpolation argument. For Algorithm 1, we take the following continuous linear interpolation between $\mu^{(t)}$ and $\mu^{(t+1)}$ over the space of probability distributions. Let $\nu_0 = \mu^{(t)}$, and

$$\nu_s = \nu_0 + \gamma s(\hat{\nu}_0 - \nu_0) = \mu^{(t)} + s\gamma(\hat{\mu}^{(t)} - \mu^{(t)}),$$

where $s \in [0, \eta]$. Then, ν_s is a linear approximation to the continuous EFP (7) and we obtain the following:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{L}(\nu_s) &= -\lambda\gamma\left(\mathrm{KL}(\nu_0\|\hat{\nu}_0) + \mathrm{KL}(\hat{\nu}_0\|\nu_0)\right)\\
&\quad + \int\lambda\gamma\log\frac{\mathrm{d}\nu_s}{\mathrm{d}\nu_0}(\theta)\cdot\frac{\mathrm{d}\hat{\nu}_0}{\mathrm{d}\hat{\nu}_s}(\theta)\mathrm{d}(\hat{\nu}_0 - \nu_0)(\theta).
\end{aligned}$$

The last term is a time-discretization error, which can be small if the step size η is small. Therefore by evaluating this error, we can extend the proof of Theorem 4.2 and obtain the reduction in objective value per iteration.

In the following, we set $t_0 = \lceil\frac{1}{\gamma\eta}\rceil$ and introduce a constant D_λ that depends on ℓ_i and λ (for complete descriptions see Appendix A). In addition, we make the following boundedness assumptions on h_i and ℓ_i. We note the upper bound of 1 can be replaced by any value.

Assumption 5.1. For any $i = 1, 2, \ldots, n$ and $z \in [1, 1]$,

$$\|h_i\|_\infty \leq 1,\ |\ell_i'(z)| \leq 1,\ 0 \leq \ell_i''(z) \leq 1.$$

The following lemma evaluates the decrease in the primal objective value in one step.

Lemma 5.2. *Let $\{\mu^{(t)}\}_{t=0}^T \subset \mathcal{P}_2$ be a sequence generated by Algorithm 1. Suppose $\gamma\eta \leq \frac{1}{8}$ holds. Then, under Assumptions 2.1 and 5.1, we get for $t \geq t_0$,*

$$\mathcal{L}(\mu^{(t+1)}) - \mathcal{L}(\mu^{(t)}) \leq -\lambda\gamma\eta\mathrm{KL}(\mu^{(t)}\|\hat{\mu}^{(t)}) + \gamma\eta^2 D_\lambda.$$

Similarly, we can evaluate the dual objective as follows.

Lemma 5.3. *Let $\{\mu^{(t)}\}_{t=0}^T \subset \mathcal{P}_2$ be a sequence generated by Algorithm 1. Suppose $\gamma\eta \leq \frac{1}{8}$ holds. Then, under Assumptions 2.1 and 5.1, we get for $t \geq t_0$,*

$$-\mathcal{D}(g_{\mu^{(t+1)}}) + \mathcal{D}(g_{\mu^{(t)}}) \leq 2\gamma\eta^2 D_\lambda.$$

Combining the above lemmas with Theorem 4.3, we have

$$\mathrm{KL}(\mu^{(t+1)}\|\hat{\mu}^{(t+1)}) \leq (1-\gamma\eta)\mathrm{KL}(\mu^{(t)}\|\hat{\mu}^{(t)}) + \frac{7\gamma\eta^2}{\lambda}D_\lambda.$$

Hence we obtain the following primal-dual convergence.

Theorem 5.4. *Let $\{\mu^{(t)}\}_{t=0}^T \subset \mathcal{P}_2$ be a sequence generated by Algorithm 1. Suppose $\gamma\eta \leq \frac{1}{8}$ holds. Then, under Assumptions 2.1 and 5.1, we get for $t \geq t_0$,*

$$\mathrm{KL}(\mu^{(t)}\|\hat{\mu}^{(t)}) \leq (1-\gamma\eta)^{t-t_0}\mathrm{KL}(\mu^{(t_0)}\|\hat{\mu}^{(t_0)}) + \frac{7\eta}{\lambda}D_\lambda.$$

6. Gradient Boosting Viewpoint

6.1. Analysis of Algorithm 1

We introduce a gradient-boosting viewpoint of the discrete-time EFP (Algorithm 1). Recall the empirical loss is given as $F_0(\mu) = \frac{1}{n}\sum_{i=1}^n \ell_i(\mathbb{E}_\mu[h_i(\theta)])$. Observe that the proximal Gibbs distribution $\hat{\mu}$ is characterized by the minimizer of the regularized linear functional:

$$\min_{\mu'} \left\{ \int \frac{\delta F_0}{\delta \mu}(\mu)(\theta) \mathrm{d}\mu'(\theta) + \lambda \mathrm{KL}(\mu' \| \nu) \right\}, \qquad (12)$$

where ν is the Gaussian distribution in proportion to $\exp(-\lambda' \|\theta\|_2^2/\lambda)$. This can be verified by computing the first variation of the objective and solving the first-order optimality condition: $\frac{\delta F_0}{\delta \mu}(\mu) + \lambda \log(\mathrm{d}\mu'/\mathrm{d}\nu) = 0$. Therefore, $\hat{\mu}$ can be interpreted as an approximation of the "negative gradient": $-\frac{\delta F_0}{\delta \mu}(\mu)$, and the EFP update (8) can be seen as a *regularized gradient boosting* method for the empirical loss F_0 over the space of probability distributions. Consequently, we can show the following convergence of $F_0(\mu^{(T)})$ based on the Frank-Wolfe argument.

Theorem 6.1. *Suppose Assumption 2.1 holds and suppose ℓ_i is L-Lipschitz smooth, $\ell_i \geq 0$, and $\|h\|_\infty \leq B$. Then, for $\{\mu^{(t)}\}_{t=0}^T \subset \mathcal{P}_2$ generate by Algorithm 1, we get*

$$F_0(\mu^{(T)}) \leq \epsilon' + (1-\eta\gamma)^T F_0(\mu^{(0)}) + \inf_{\xi \in \mathcal{P}_2} \mathcal{L}(\xi),$$

where $\epsilon' = 2\eta\gamma B^2 L$.

Based on this theorem, we can compute the required time-complexity to achieve an ϵ-error for a given precision $\epsilon > 0$ in the following sense: $F_0(\mu^{(T)}) \leq \inf_{\xi \in \mathcal{P}_2} \mathcal{L}(\xi) + \epsilon$. Concretely, by setting $\eta\gamma = \epsilon/(4B^2L)$, the complexity is

$$T = \frac{4B^2L}{\epsilon} \log\left(\frac{2F_0(\mu_0)}{\epsilon}\right).$$

Remark 6.2. The infimum of $\mathcal{L}$ can be upper bounded as follows: for any $r > 0$, $\inf_{\xi \in \mathcal{P}_2} \mathcal{L}(\xi) \leq \inf_{\xi \in \mathbb{B}_r} F_0(\xi) + \lambda r$, where $\mathbb{B}_r = \{\xi \in \mathcal{P}_2 \mid \mathrm{KL}(\xi \| \nu) \leq r\}$. Therefore, by setting $\lambda = \epsilon r^{-1}$ the above result implies the convergence to an approximate solution of F_0 over $\mathbb{B}_r$, that is, $F_0(\mu^{(T)}) \leq \inf_{\xi \in \mathbb{B}_r} F_0(\xi) + 2\epsilon$. However, we note that the small λ could potentially affect the complexity of sampling steps for $\hat{\mu}^{(t)}$ in the practical implementation such as Algorithm 2.

6.2. Analysis of Algorithm 2

Next, we outline the extension of Theorem 6.1 that applies to Algorithm 2 (i.e., the finite-particle implementation). Specifically, Theorem 6.1 can be relaxed so that EFP allows for a tolerance factor in the next iteration $\mu^{(t+1)}$; that is, for a given $\rho \geq 0$,

$$F_0(\mu^{(t+1)}) \leq \eta\gamma\rho + F_0((1-\eta\gamma)\mu^{(t)} + \eta\gamma\hat{\mu}^{(t)}). \qquad (13)$$

For the convenience of explanation in this section, we denote by $\mu'^{(t+1)}$ the exact update $(1-\eta\gamma)\mu^{(t)} + \eta\gamma\hat{\mu}^{(t)}$ in Algorithm 1. Note that $\mu'^{(t+1)}$ satisfies the above condition with $\rho = 0$. The above relaxation (13) yields a guarantee similar to Theorem 6.1 with $\epsilon' = \rho + 2\eta\gamma B^2 L$. Let us apply this extended version of Theorem 6.1 to the case of $\mu^{(t+1)} = (1-\eta\gamma)\mu^{(t)} + \eta\gamma\nu^{(t)}$, where $\nu^{(t)} = \frac{1}{m}\sum_{r=1}^m \delta_{\theta_r^{(t)}}$ is an empirical distribution attained by Langevin Monte Carlo for $\hat{\mu}^{(t)}$. This is to say, we modify Algorithm 1 by replacing $\hat{\mu}^{(t)}$ with an empirical distribution $\nu^{(t)}$. If $\|h_i\|_\infty \leq B$ and ℓ_i is C-Lipschitz continuous, Pinsker's inequality leads to the following upper-bound on $|F_0(\mu'^{(t+1)}) - F_0(\mu^{(t+1)})|$:

$$\eta\gamma C B \sqrt{2\mathrm{KL}(\overline{\nu}^{(t)} \| \hat{\mu}^{(t)})} + \frac{\eta\gamma C}{n} \sum_{i=1}^n |\mathbb{E}_{\overline{\nu}^{(t)}}[h_i] - \mathbb{E}_{\nu^{(t)}}[h_i]|,$$

where we set $\overline{\nu}^{(t)} = \mathrm{Law}(\theta_r^{(t)})$. Therefore, we can estimate the particle and iteration complexities to satisfy the required precision (13) by applying the convergence rate of the unadjusted Langevin Monte Carlo (Vempala & Wibisono, 2019) together with a standard concentration inequality as done in Nitanda et al. (2021); Oko et al. (2022).

Theorem 6.3. *Suppose Assumption 2.1 holds, ℓ_i is C-Lipschitz continuous and L-Lipschitz smooth, $\ell_i \geq 0$, and $\|h\|_\infty \leq B$, and ∇h_i is D-Lipschitz continuous regarding $\|\cdot\|_2$. Suppose the underlying Langevin Monte Carlo is run with the initialization $\overline{\theta}_r^{(0)} \sim \mu^{(0)}$. Given $\delta \in (0,1)$, hyperparameters η', S, m of Algorithm 2, are set as follows: $\eta' \leq \min\left\{\frac{\alpha}{4L'^2}, \frac{\alpha\rho^2}{128dB^2C^2L'^2}\right\}$, $S \geq \frac{1}{\alpha\eta'} \log \frac{\alpha \mathrm{KL}(\mu^{(0)} \| \hat{\mu}^{(t)})}{8\eta' d L'^2}$, $m \geq \frac{8B^2C^2}{\rho^2} \log \frac{2nT}{\delta}$, where α is the LSI-constant of $\hat{\mu}^{(t)}$ and $L' = \frac{1}{\lambda}(CD + 2\lambda')$. Then, for $\{\mu^{(t)}\}_{t=0}^T \subset \mathcal{P}_2$ generate by Algorithm 2, we get with probability at least $1-\delta$,*

$$F_0(\mu^{(T)}) \leq \epsilon' + (1-\eta\gamma)^T F_0(\mu^{(0)}) + \inf_{\xi \in \mathcal{P}_2} \mathcal{L}(\xi),$$

where $\epsilon' = \rho + 2\eta\gamma B^2 L$.

Theorem 6.3 entails the particle and iteration complexity to achieve a given ϵ-accurate solution in the following sense: $F_0(\mu^{(T)}) \leq \inf_{\xi \in \mathcal{P}_2} \mathcal{L}(\xi) + \epsilon$. In particular, using the evaluation (Holley & Stroock, 1987) (see Appendix B), $\hat{\mu}^{(t)}$ satisfies LSI with at least the constant $\alpha = \frac{2\lambda'}{\lambda} \exp\left(-\frac{4BC}{\lambda}\right)$. We set $\rho = \Theta(\epsilon)$, $\eta\gamma = \Theta(\epsilon)$, and $T = \Theta\left(\epsilon^{-1} \log \frac{F_0(\mu^{(0)})}{\epsilon}\right) = \tilde{\Theta}(\epsilon^{-1})$. For the step size η', it is sufficient to consider the case of $\eta' = \Theta\left(\alpha\lambda^2\epsilon^2\right)$. Then, $S = \Theta\left(\alpha^{-2}\lambda^{-2}\epsilon^{-2} \log \frac{\mathrm{KL}(\mu^{(0)} \| \hat{\mu}^{(t)})}{8\epsilon^2 d L^2}\right) = \tilde{\Theta}(\alpha^{-2}\lambda^{-2}\epsilon^{-2})$ and $m = \Theta\left(\epsilon^{-2} \log \frac{2nT}{\delta}\right) = \tilde{\Theta}(\epsilon^{-2})$, where the notation $\tilde{\Theta}$ hides the logarithmic term for simplicity. Therefore, particle and iteration complexities are as follows, respectively:

$$m = \tilde{\Theta}(\epsilon^{-2}), \; ST = \tilde{\Theta}(\alpha^{-2}\lambda^{-2}\epsilon^{-3}).$$

7. Experiments

Note that our theory also guarantees the convergence of the proximal Gibbs distribution $\hat{\underline{\mu}}^{(T-1)}$ to the (regularized) global optimal solution. Therefore, in the following, we evaluate the performance of $\hat{\underline{\mu}}^{(T-1)}$ as well as $\underline{\mu}^{(T)}$ in the applications of training two-layer neural networks as well as image synthesis using transparent triangles.

7.1. Two-layer Neural Networks

As previously remarked, optimizing a two-layer neural network in the mean-field regime is one important application of the EFP algorithm. We write $h(\theta, x) = \sigma(\theta^\top x)$ as one neuron with trainable parameters $\theta \in \mathbb{R}^d$. In this setting, the mean-field model $\mathbb{E}_\mu[h(\theta, \cdot)]$ is an infinite-width neural network, and $\frac{1}{m}\sum_{r=1}^m h(\theta_r, \cdot)$ is an M-particle approximation of the infinite-width limit (i.e., a neural network with m hidden neurons).

We consider a student-teacher setting where the labels are generated by a teacher network with cosine activation; we set $n = 500, d = 5$. The student model is a two-layer network with tanh activation and width $m = 1000$. The training objective is to minimize the empirical squared error, and the regularization hyperparameters are set to $\lambda = \lambda' = 0.01$. We optimize the neural network using EFP with an outer-loop step size $\eta\gamma = 0.01$. At each iteration, we approximate the proximal Gibbs measure $\hat{\mu}_t$ via the Langevin Monte Carlo algorithm with step size $\eta' = 0.01$.

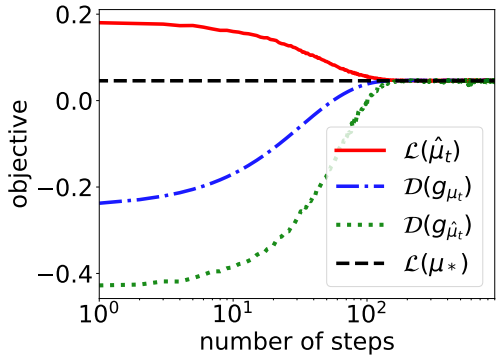

Figure 2. Illustration of primal-dual convergence: learning two-layer neural net with EFP ($\lambda = \lambda' = 10^{-2}$).

In Figure 2 we report the primal and dual objective values, in which the entropy term is computed by the k-nearest neighbors estimator (Kozachenko & Leonenko, 1987). The optimal value $\mathcal{L}(\mu_*)$ is approximated using the particle dual averaging (PDA) algorithm (Nitanda et al., 2021) which also globally optimizes the objective (6). Observe that the duality gap vanishes as predicted by our theoretical analysis.

7.2. Image Synthesis with Triangles

We also consider the image synthesis experiment in Tian & Ha (2022), where the goal is to "approximate" an image via integrating transparent triangles. To draw the triangles, we utilize a differentiable render (Laine et al., 2020) that provides a differentiable map to translate the parameters θ (representing the colors and vertexes) into a transparent triangle $h(\theta)$. Then, we can construct an image $\mathbb{E}_{\theta\sim\mu}[h(\theta)]$ by integrating $h(\theta)$ with the probability distribution μ. For simple explanation, we take a $W \times H$ image of a single channel, i.e., $h(\theta) = \{h_{ij}(\theta)\}_{i,j} \in \mathbb{R}^{WH}$. We formalize the problem of approximating a given image $J \in \mathbb{R}^{WH}$ as a regularized regression problem using the squared error:

$$\frac{1}{WH}\sum_{i,j}\ell_{i,j}(\mathbb{E}_\mu[h_{ij}(\theta)]) = \frac{1}{WH}\sum_{i,j}(J_{i,j} - \mathbb{E}_\mu[h_{ij}(\theta)])^2.$$

Then, we can apply Algorithm 2 to minimize this function with regularization. We adopt the image "Mona Lisa" of size 256×256 as the target. See Figure 3 for the original target image and one generated image $H^{(T)} = \{H^{(t)}_{i,j}\}_{i,j}$ using EFP (Algorithm 2).

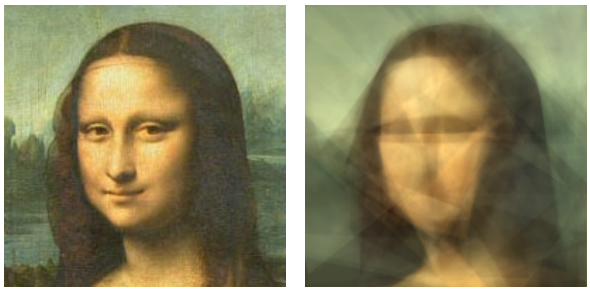

Figure 3. We run Algorithm 2 with $\lambda = 10^{-5}$, $\lambda' = 10^{-4}$, $T = 2000$, $S = 10$, $m = 1000$, $\eta \cdot \gamma = 0.01$ to fit the target image. As for the step size for Langevin Monte Carlo, we used cosine annealing from 0.1 to 0.01. The figure depicts the target image "Mona Lisa" (left) and an obtained image (right).

In Figure 4 we compare the generated images $H^{(T)}$ with $\frac{1}{m}\sum_{r=1}^m h(\theta_r^{(T-1)})$ using different number of triangles of $m = 200$ and $m = 1000$; we also report the generated image under both $\underline{\mu}^{(T)}$ and the Gibbs distribution, $\hat{\underline{\mu}}^{(T-1)}$ as explained in Section 3.2. Figure 4 depicts these generated images. Finally, we plot the squared error of each image during optimization in Figure 5.

We observe that EFP is capable of generating reasonable images. Moreover, as the number of particle m increases, the quality of images generated by both $\underline{\mu}^{(T)}$ and $\hat{\underline{\mu}}^{(T-1)}$ improve and the discrepancy between them becomes smaller. This is consistent with our theory, which guarantees the convergence of both distributions to the optimal solution when m becomes large.

Figure 4. Comparison of images generated by $\underline{\mu}^{(T)}$ and $\hat{\underline{\mu}}^{(T-1)}$. We set $m = 200$ for the first two figures on the left, and $m = 1000$ for the two on the right. In both cases, the left and right figures correspond to images by $\underline{\mu}^{(T)}$ and $\hat{\underline{\mu}}^{(T-1)}$, respectively.

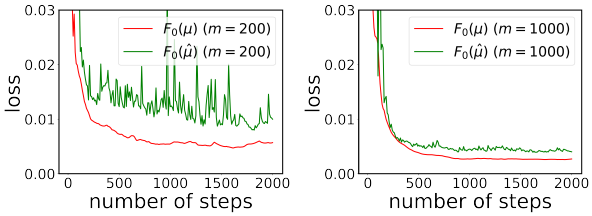

Figure 5. Squared error achieved by $\underline{\mu}^{(T)}$ and $\hat{\underline{\mu}}^{(T-1)}$ with the number of triangles (particle size) $m = 200, 1000$.

Conclusion

In this paper, we presented a concise primal-dual analysis of the entropic fictitious play (EFP) algorithm for the finite-sum minimization problem. Specifically, we showed the convergence in both continuous- and discrete-time in terms of the primal and dual objectives, which is equivalent to the convergence of the KL divergence between μ and $\hat{\mu}$. Our analysis also leads to an efficient implementation of EFP which resolves the huge memory consumption issue. Moreover, we introduced a gradient-boosting perspective of EFP which gives a new discrete-time global convergence analysis; under this perspective, EFP can be regarded as the memory-efficient gradient-boosting method performing on the space of probability distributions. Finally, we empirically verified our theory in the application of training two-layer neural networks and an image synthesis using transparent triangles. An interesting future work is the generalization analysis for mean-field models. Especially, learnability in the classification setting is of interest for the combination of our convergence result and margin theory.

Acknowledgment

AN was partially supported by JSPS Kakenhi (22H03650). KO was partially supported by Fujitsu Ltd. DW was partially supported by a Borealis AI Fellowship. TS was partially supported by JSPS KAKENHI (20H00576) and JST CREST.

References

Abbe, E., Adsera, E. B., and Misiakiewicz, T. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pp. 4782–4887. PMLR, 2022.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems 35*, 2022.

Bach, F. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

Bakry, D. and Émery, M. Diffusions hypercontractives in sem. probab. xix lnm 1123, 1985.

Bauschke, H. H., Combettes, P. L., et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Brown, G. W. Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation*, 13(1): 374–376, 1951.

Cardaliaguet, P. and Hadikhanloo, S. Learning in mean field games: the fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(2):569–591, 2017.

Chen, F., Ren, Z., and Wang, S. Uniform-in-time propagation of chaos for mean field langevin dynamics. *arXiv preprint arXiv:2212.03050*, 2022.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. Stein points. In *International Conference on Machine Learning*, pp. 844–853. PMLR, 2018.

Chizat, L. Mean-field langevin dynamics: Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems 31*, pp. 3040–3050, 2018.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems 32*, pp. 9111–9121, 2019.

Hadikhanloo, S. and Silva, F. J. Finite mean field games: fictitious play and convergence to a first order continuous mean field game. *Journal de Mathématiques Pures et Appliquées*, 132:369–397, 2019.

Holley, R. and Stroock, D. Logarithmic sobolev inequalities and stochastic ising models. *Journal of statistical physics*, 46(5-6):1159–1194, 1987.

Hu, K., Ren, Z., Siska, D., and Szpruch, L. Mean-field langevin dynamics and energy landscape of neural networks. *arXiv preprint arXiv:1905.07769*, 2019.

Kozachenko, L. and Leonenko, N. N. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., and Aila, T. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020.

Lavigne, P. and Pfeiffer, L. Generalized conditional gradient and learning in potential mean field games. *arXiv preprint arXiv:2209.12772*, 2022.

Li, Y., Ma, T., and Zhang, H. R. Learning over-parametrized two-layer neural networks beyond ntk. In *Proceedings of Conference on Learning Theory 33*, pp. 2613–2682, 2020.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Nishikawa, N., Suzuki, T., Nitanda, A., and Wu, D. Two-layer neural network on infinite dimensional data: global optimization guarantee in the mean-field regime. In *Advances in Neural Information Processing Systems 35*, 2022.

Nitanda, A. and Suzuki, T. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

Nitanda, A., Wu, D., and Suzuki, T. Particle dual averaging: Optimization of mean field neural networks with global convergence rate analysis. In *Advances in Neural Information Processing Systems 34*, pp. 19608–19621, 2021.

Nitanda, A., Wu, D., and Suzuki, T. Convex analysis of the mean field langevin dynamics. In *Proceedings of International Conference on Artificial Intelligence and Statistics 25*, pp. 9741–9757, 2022.

Oko, K., Suzuki, T., Nitanda, A., and Wu, D. Particle stochastic dual coordinate ascent: Exponential convergent algorithm for mean field neural network optimization. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Perrin, S., Pérolat, J., Laurière, M., Geist, M., Elie, R., and Pietquin, O. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems*, 33:13199–13213, 2020.

Ren, Z. and Wang, S. Entropic fictitious play for mean field optimization problem. *arXiv preprint arXiv:2202.05841*, 2022.

Rockafellar, R. T. *Convex Analysis*. Princeton University Press, Princeton, 1970.

Rotskoff, G. M. and Vanden-Eijnden, E. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems 31*, pp. 7146–7155, 2018.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

Suzuki, T., Nitanda, A., and Wu, D. Uniform-in-time propagation of chaos for the mean field gradient langevin dynamics. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Tian, Y. and Ha, D. Modern evolution strategies for creativity: Fitting concrete images and abstract concepts. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, pp. 275–291, 2022.

Vempala, S. and Wibisono, A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems 32*, pp. 8094–8106, 2019.

A. Analysis in discrete-time

Lemma A.1. *Suppose Assumption 5.1 holds. Then, any two probability distributions μ_1, μ_2 satisfy*

$$C_\lambda^{-1} \le \frac{\hat{\mu}_1(\theta)}{\hat{\mu}_2(\theta)} \le C_\lambda, \tag{14}$$

with $C_\lambda := \exp(4/\lambda)$.

Moreover, let μ_1, μ_2 be two probability distributions, and $\mu := (1-t)\mu_1 + t\mu_2$ be an interpolation with $t \in [0, \frac{1}{2}]$. Then, we have that

$$-4t \le \frac{\hat{\mu}(\theta)}{\hat{\mu}_1(\theta)} - 1 \le 4tC_\lambda. \tag{15}$$

Proof. Because of $\|h_i\|_\infty \le 1$ and $|(g_\mu)_i| = |\ell_i'(x)| \le 1$, the conclusion follows immediately by the following formulation:

$$\frac{\hat{\mu}_1(\theta)}{\hat{\mu}_2(\theta)} = \frac{\exp\left(-\frac{1}{\lambda}\left(\frac{1}{n}\sum_{i=1}^n h_i(\theta) g_{\mu_1,i} + \lambda'\|\theta\|_2^2\right)\right)}{\exp\left(-\frac{1}{\lambda}\left(\frac{1}{n}\sum_{i=1}^n h_i(\theta) g_{\mu_2,i} + \lambda'\|\theta\|_2^2\right)\right)} \frac{\int \exp\left(-\frac{1}{\lambda}\left(\frac{1}{n}\sum_{i=1}^n h_i(\theta') g_{\mu_2,i} + \lambda'\|\theta'\|_2^2\right) \mathrm{d}\theta'\right)}{\int \exp\left(-\frac{1}{\lambda}\left(\frac{1}{n}\sum_{i=1}^n h_i(\theta'') g_{\mu_1,i} + \lambda'\|\theta''\|_2^2\right) \mathrm{d}\theta''\right)}.$$

In the same way, to bound $\frac{\hat{\mu}(\theta)}{\hat{\mu}_1(\theta)}$, we consider the density ratio:

$$\frac{\hat{\mu}(\theta)}{\hat{\mu}_1(\theta)} = \frac{\exp\left(-\frac{1}{\lambda}\left(\frac{1}{n}\sum_{i=1}^n h_i(\theta) g_{\mu,i} + \lambda'\|\theta\|_2^2\right)\right)}{\exp\left(-\frac{1}{\lambda}\left(\frac{1}{n}\sum_{i=1}^n h_i(\theta) g_{\mu_1,i} + \lambda'\|\theta\|_2^2\right)\right)} \frac{\int \exp\left(-\frac{1}{\lambda}\left(\frac{1}{n}\sum_{i=1}^n h_i(\theta') g_{\mu_1,i} + \lambda'\|\theta'\|_2^2\right) \mathrm{d}\theta'\right)}{\int \exp\left(-\frac{1}{\lambda}\left(\frac{1}{n}\sum_{i=1}^n h_i(\theta'') g_{\mu,i} + \lambda'\|\theta''\|_2^2\right) \mathrm{d}\theta''\right)}. \tag{16}$$

Note that

$$\begin{aligned}
|g_{\mu,i} - g_{\mu_1,i}| &= \left|\ell_i'\left(\int h_i(\theta)\mathrm{d}\mu(\theta)\right) - \ell_i'\left(\int h_i(\theta)\mathrm{d}\mu_1(\theta)\right)\right| \\
&\le \left|\int h_i(\theta)\mathrm{d}\mu(\theta) - \int h_i(\theta)\mathrm{d}\mu_1(\theta)\right| \\
&= t\left|\int h_i(\theta)\mathrm{d}\mu_2(\theta) - \int h_i(\theta)\mathrm{d}\mu_1(\theta)\right| \\
&\le 2t.
\end{aligned}$$

Therefore, the equation (16) is upper and lower bounded by $\exp(4t/\lambda)$ and $\exp(-4t/\lambda)$, respectively. Note that we have $\exp(4t/\lambda) \le 1 + 4t\exp(4/\lambda)$ and $\exp(-4t/\lambda) \ge 1 - 4t$ if $0 \le t \le \frac{1}{4}$ holds, which concludes the proof. □

Lemma A.2. *Suppose Assumption 5.1 and $\gamma\eta < 1$ hold. Then, for $t \ge t_0 := \lceil \frac{1}{\gamma\eta} \rceil$, we get $\frac{1}{2C_\lambda}\hat{\mu}^{(t)} \le \mu^{(t)}$.*

Proof. First, we observe that

$$\begin{aligned}
\mu^{(t)} &= (1-\gamma\eta)\mu^{(t-1)} + \eta\gamma\hat{\mu}^{(t-1)} \\
&= \cdots = (1-\gamma\eta)^t\mu^{(0)} + \eta\gamma\sum_{s=0}^{t-1}(1-\gamma\eta)^{t-s-1}\hat{\mu}^{(s)}.
\end{aligned} \tag{17}$$

According to the bounds (14) of Lemma A.1, $\hat{\mu}^s(\theta) \ge C_\lambda^{-1}\hat{\mu}^{(t)}$ holds for each s and θ. Therefore, (17) yields that

$$\mu^{(t)} \ge \eta\gamma\sum_{s=0}^{t-1}(1-\gamma\eta)^{t-s-1}C_\lambda^{-1}\hat{\mu}^{(t)} \ge (1-(1-\gamma\eta)^t)C_\lambda^{-1}\hat{\mu}^{(t)}.$$

By letting $t \ge t_0 := \lceil \frac{1}{\gamma\eta} \rceil \ge \log_{1-\gamma\eta}\frac{1}{2}$, we obtain that $\mu^{(t)} \ge \frac{1}{2C_\lambda}\hat{\mu}^{(t)}$.

□

We then present the analysis for each iteration of Algorithm 1 for the primal and dual objectives.

Lemma A.3. *Suppose Assumption 5.1 and $\gamma\eta \le \frac{1}{8}$ hold. Then, for $t \ge t_0 := \lceil \frac{1}{\gamma\eta} \rceil$, we get*

$$\mathcal{L}(\mu^{(t+1)}) - \mathcal{L}(\mu^{(t)}) \le -\lambda\gamma\eta \mathrm{KL}(\mu^{(t)} \| \hat{\mu}^{(t)}) + \gamma\eta^2(3C_\lambda + 5).$$

Proof. We consider the following continuous interpolation between $\mu^{(t)}$ and $\mu^{(t+1)}$. Let $\nu_0 = \mu^{(t)}$, and $\nu_\eta = \nu_0 + \gamma\eta(\hat{\nu}_0 - \nu_0) = \mu^{(t)} + \eta\gamma(\hat{\mu}^{(t)} - \mu^{(t)})$. We then revisit the analysis on primal convergence (Theorem 4.2).

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\eta}\mathcal{L}(\nu_\eta) &= \int \frac{\delta\mathcal{L}}{\delta\mu}(\nu_\eta)(\theta)\frac{\mathrm{d}\nu_\eta}{\mathrm{d}\eta}(\theta) \\
&= \int \lambda\gamma \log\frac{\mathrm{d}\nu_\eta}{\mathrm{d}\hat{\nu}_\eta}(\theta)\mathrm{d}(\hat{\nu}_0 - \nu_0)(\theta) \\
&= \int \lambda\gamma \log\frac{\mathrm{d}\nu_0}{\mathrm{d}\hat{\nu}_0}(\theta)\mathrm{d}(\hat{\nu}_0 - \nu_0)(\theta) + \int \lambda\gamma \log\frac{\mathrm{d}\nu_\eta}{\mathrm{d}\nu_0}(\theta)\mathrm{d}(\hat{\nu}_0 - \nu_0)(\theta) + \int \lambda\gamma \log\frac{\mathrm{d}\hat{\nu}_0}{\mathrm{d}\hat{\nu}_\eta}(\theta)\mathrm{d}(\hat{\nu}_0 - \nu_0)(\theta) \quad (18)
\end{aligned}$$

The first term is equal to $-\lambda\gamma\left(\mathrm{KL}(\nu_0\|\hat{\nu}_0) + \mathrm{KL}(\hat{\nu}_0\|\nu_0)\right)$ as previously.

To bound the second term, we bound the density ratio $\frac{\mathrm{d}\nu_\eta}{\mathrm{d}\nu_0}(\theta)$ as

$$\frac{\mathrm{d}\nu_\eta}{\mathrm{d}\nu_0}(\theta) = \frac{\mathrm{d}(\nu_0 - \gamma\eta(\nu_0 - \hat{\nu}_0))}{\mathrm{d}\nu_0}(\theta) = \frac{(1-\gamma\eta)\mathrm{d}\mu^{(t)}}{\mathrm{d}\mu^{(t)}}(\theta) + \gamma\eta\frac{\mathrm{d}\hat{\mu}^{(t)}}{\mathrm{d}\mu^{(t)}}(\theta) \begin{cases} \ge 1 - \gamma\eta, \\ \le (1-\gamma\eta) + 2C_\lambda\gamma\eta = 1 + \gamma\eta(2C_\lambda - 1), \end{cases}$$

where we used Lemma A.2 the upper bound. Therefore, we have that

$$\int \lambda\gamma \log\frac{\mathrm{d}\nu_\eta}{\mathrm{d}\nu_0}(\theta)\mathrm{d}(\hat{\nu}_0 - \nu_0)(\theta) \le \log(1 + \gamma\eta(2C_\lambda - 1)) - \log(1 - \gamma\eta) \le \gamma\eta(2C_\lambda + 2),$$

where we used that $\log(1+x) \le x$ and $\log(1-x) \ge 2x$ $(0 \le x \le \frac{1}{2})$ for the last inequality.

The third term is bounded in a similar manner. The bounds (15) from Lemma A.1 yield that

$$\int \lambda\gamma \log\frac{\mathrm{d}\hat{\nu}_0}{\mathrm{d}\hat{\nu}_\eta}(\theta)\mathrm{d}(\hat{\nu}_0 - \nu_0)(\theta) \le \log(1 + 4\gamma\eta C_\lambda) - \log(1 - 4\gamma\eta).$$

RHS is further bounded by $4\gamma\eta(C_\lambda + 2)$, where we used the fact that $\log(1+x) \le x$ and $\log(1-x) \ge 2x$ $(0 \le x \le \frac{1}{2})$.

Putting it all together, (18) is bounded by

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\mathcal{L}(\nu_\eta) \le -\lambda\gamma\left(\mathrm{KL}(\nu_0\|\hat{\nu}_0) + \mathrm{KL}(\hat{\nu}_0\|\nu_0)\right) + \gamma\eta(6C_\lambda + 10) \le -\lambda\gamma\mathrm{KL}(\mu^{(t)}\|\hat{\mu}^{(t)}) + \gamma\eta(6C_\lambda + 10).$$

By integrating LHS and RHS, we obtain that

$$\mathcal{L}(\nu_\eta) - \mathcal{L}(\nu_0) \le -\lambda\gamma\eta\mathrm{KL}(\mu^{(t)}\|\hat{\mu}^{(t)}) + \gamma\eta^2(3C_\lambda + 5),$$

and $\mathcal{L}(\mu^{(t+1)}) - \mathcal{L}(\mu^{(t)}) = \mathcal{L}(\nu_\eta) - \mathcal{L}(\nu_0)$ concludes the proof. □

Lemma A.4. *Suppose Assumption 5.1 and $\gamma\eta \le \frac{1}{8}$ hold. Then, for $t \ge t_0 := \lceil \frac{1}{\gamma\eta} \rceil$, we get*

$$-\mathcal{D}(g_{\mu^{(t+1)}}) + \mathcal{D}(g_{\mu^{(t)}}) \le 2\gamma\eta^2(1 + 2C_\lambda).$$

Proof. Let ν_η be defined in the same way as previously. We consider the evolution of $\frac{\mathrm{d}}{\mathrm{d}\eta}\mathcal{D}(g_{\nu_\eta})$ based on the analysis for Theorem 4.4. We have that

$$\begin{aligned}
-\frac{\mathrm{d}}{\mathrm{d}\eta}\mathcal{D}(g_{\nu_\eta}) &= -\nabla\mathcal{D}(g_{\nu_\eta})^\top \frac{\mathrm{d}g_{\nu_\eta}}{\mathrm{d}\nu} \\
&= \frac{1}{n}\sum_{i=1}^n \left(\ell_i^{*\prime}(g_{\nu_\eta,i}) - \frac{1}{\int q_{g_{\nu_\eta}}(\theta)\mathrm{d}\theta}\int h_i(\theta)q_{g_{\nu_\eta}}(\theta)\mathrm{d}\theta\right) \cdot \frac{\mathrm{d}}{\mathrm{d}\eta}\ell_i'(\mathbb{E}_{\nu_\eta}[h_i(\theta)]) \\
&= -\frac{\gamma}{n}\sum_{i=1}^n (\mathbb{E}_{\nu_\eta}[h_i(\theta)] - \mathbb{E}_{\hat{\nu}_\eta}[h_i(\theta)])(\mathbb{E}_{\nu_0}[h_i(\theta)] - \mathbb{E}_{\hat{\nu}_0}[h_i(\theta)])\ell_i''(\mathbb{E}_{\nu_\eta}[h_i]). \quad (19)
\end{aligned}$$

where we used $\hat{\nu}_\eta \propto q_{g_{\nu_\eta}}(\theta)\mathrm{d}\theta$, $\ell_i^{*\prime} = (\ell_i')^{-1}$, and

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\ell_i'(\mathbb{E}_{\nu_\eta}[h_i(\theta)]) = \ell_i''(\mathbb{E}_{\nu_\eta}[h_i]) \int h_i(\theta)\frac{\mathrm{d}\nu_\eta}{\mathrm{d}\eta}(\theta) = \ell_i''(\mathbb{E}_{\nu_\eta}[h_i])(\mathbb{E}_{\hat{\nu}_0}[h_i] - \mathbb{E}_{\nu_0}[h_i]).$$

We bound the difference between $(\mathbb{E}_{\nu_\eta}[h_i(\theta)] - \mathbb{E}_{\hat{\nu}_\eta}[h_i(\theta)])$ and $(\mathbb{E}_{\nu_0}[h_i(\theta)] - \mathbb{E}_{\hat{\nu}_0}[h_i(\theta)])$. First, $\mathbb{E}_{\nu_\eta}[h_i(\theta)]$ is evaluated as

$$\begin{aligned}&\mathbb{E}_{\nu_\eta}[h_i(\theta)] = (1-\gamma\eta)\mathbb{E}_{\nu_0}[h_i(\theta)] + \gamma\eta\mathbb{E}_{\hat{\nu}_0}[h_i(\theta)]\\ &\therefore |\mathbb{E}_{\nu_\eta}[h_i(\theta)] - \mathbb{E}_{\nu_0}[h_i(\theta)]| \le 2\gamma\eta,\end{aligned} \tag{20}$$

where we used $\|h_i\|_\infty \le 1$. Next, according to (15) of Lemma A.1, we have that

$$|\mathbb{E}_{\hat{\nu}_\eta}[h_i(\theta)] - \mathbb{E}_{\hat{\nu}_0}[h_i(\theta)]| = \left|\int h_i \mathrm{d}(\hat{\nu}_\eta - \hat{\nu}_0)\right| \le 4\gamma\eta C_\lambda \left|\int h_i \mathrm{d}\hat{\nu}_0\right| \le 4\gamma\eta C_\lambda. \tag{21}$$

Now applying (20) and (21) to (19), we get

$$-\frac{\mathrm{d}}{\mathrm{d}\eta}\mathcal{D}(g_{\nu_\eta}) = -\frac{\gamma}{n}\sum_{i=1}^n (\mathbb{E}_{\nu_0}[h_i(\theta)] - \mathbb{E}_{\hat{\nu}_0}[h_i(\theta)])\ell_i''(\mathbb{E}_{\nu_\eta}[h_i])^2 + 4\gamma\eta(1+2C_\lambda)L \le 4\gamma\eta(1+2C_\lambda).$$

where we also used $\|h_i\|_\infty \le 1$ and $0 \le \ell_i'' \le 1$. By integrating LHS and RHS, we obtain that

$$-\mathcal{D}(g_{\nu_\eta}) + \mathcal{D}(g_{\nu_0}) \le 2\gamma\eta^2(1+2C_\lambda).$$

$-\mathcal{D}(g_{\mu^{(t+1)}}) + \mathcal{D}(g_{\mu^{(t)}}) = -\mathcal{D}(g_{\nu_\eta}) + \mathcal{D}(g_{\nu_0})$ by definition yields the assertion. □

Theorem A.5. *Suppose Assumption 5.1 and $\gamma\eta \le \frac{1}{8}$ hold. Then, for $t \ge t_0 := \lceil \frac{1}{\gamma\eta} \rceil$, we get*

$$\mathrm{KL}(\mu^{(t)}\|\hat{\mu}^{(t)}) \le (1-\gamma\eta)^{t-t_0}\mathrm{KL}(\mu^{(t_0)}\|\hat{\mu}^{(t_0)}) + \frac{7\eta}{\lambda}(1+C_\lambda)$$

for all $t \ge t_0$, where $t_0 = \lceil \frac{1}{\gamma\eta} \rceil$ and $C_\lambda = \exp(4/\lambda)$.

Proof. According to Lemmas A.3 and A.4, we have that

$$\mathcal{L}(\mu^{(t+1)}) - \mathcal{D}(g_{\mu^{(t+1)}}) - (\mathcal{L}(\mu^{(t)}) - \mathcal{D}(g_{\mu^{(t+1)}})) \le -\lambda\gamma\eta\mathrm{KL}(\mu^{(t)}\|\hat{\mu}^{(t)}) + 7\gamma\eta^2(1+C_\lambda).$$

LHS is equal to $\lambda\mathrm{KL}(\mu^{(t+1)}\|\hat{\mu}^{(t+1)}) - \lambda\mathrm{KL}(\mu^{(t)}\|\hat{\mu}^{(t)})$, by using Theorem 4.3. Thus, we obtain that

$$\mathrm{KL}(\mu^{(t+1)}\|\hat{\mu}^{(t+1)}) - \mathrm{KL}(\mu^{(t)}\|\hat{\mu}^{(t)}) \le -\gamma\eta\mathrm{KL}(\mu^{(t)}\|\hat{\mu}^{(t)}) + \frac{7\gamma\eta^2}{\lambda}(1+C_\lambda), \tag{22}$$

and Grönwall's inequality yields that

$$\left(\mathrm{KL}(\mu^{(t)}\|\hat{\mu}^{(t)}) - \frac{\eta}{\lambda}(7+7C_\lambda)\right) \le (1-\gamma\eta)^{t-t_0}\left(\mathrm{KL}(\mu^{(t_0)}\|\hat{\mu}^{(t_0)}) - \frac{7\eta}{\lambda}(1+C_\lambda)\right). \tag{23}$$

Therefore, we obtain that

$$\mathrm{KL}(\mu^{(t)}\|\hat{\mu}^{(t)}) \le (1-\gamma\eta)^{t-t_0}\mathrm{KL}(\mu^{(t_0)}\|\hat{\mu}^{(t_0)}) + \frac{7\eta}{\lambda}(1+C_\lambda)$$

for all $t \ge t_0$. □

Corollary A.6. *In order to make the dualty gap smaller than ε, i.e. $\mathcal{L}(\mu^{(t)}) - \mathcal{D}(g_{\mu^{(t)}}) = \lambda\mathrm{KL}(\mu^{(t)}\|\hat{\mu}^{(t)}) \le \varepsilon$, it suffices to take $\eta = \frac{\varepsilon}{14(1+C_\lambda)}$ and $t \ge 14\gamma(1+C_\lambda)\varepsilon^{-1}\log(2e\lambda\varepsilon^{-1}\mathrm{KL}(\mu^{(t_0)}\|\hat{\mu}^{(t_0)}))$.*

B. Gradient Boosting Viewpoint

B.1. Convergence analysis of an inexact EFP

We here relax the next iteration $\mu^{(t+1)}$ to an inexact version. That is, we replace $\mu^{(t+1)} = (1-\eta\gamma)\mu^{(t)} + \eta\gamma\hat{\mu}^{(t)}$ in Algorithm 1 with a probability distribution that satisfies

$$F_0(\mu^{(t+1)}) \le \eta\gamma\rho + F_0((1-\eta\gamma)\mu^{(t)} + \eta\gamma\hat{\mu}^{(t)}). \tag{24}$$

Then, we prove the generalized version of Theorem 6.1 which can apply to Algorithm 2 with the above modification.

Theorem B.1. *Suppose Assumption 2.1 holds and suppose ℓ_i is L-Lipschitz smooth, $\ell_i \ge 0$, and $\|h\|_\infty \le B$. Then, for $\{\mu^{(t)}\}_{t=0}^T \subset \mathcal{P}_2$ defined above, we get*

$$F_0(\mu^{(T)}) \le \epsilon' + (1-\eta\gamma)^T F_0(\mu^{(0)}) + \inf_{\xi\in\mathcal{P}_2} \mathcal{L}(\xi),$$

where $\epsilon' = \rho + 2\eta\gamma B^2 L$.

Proof. For notational simplicity, we denote $h_{i,\mu} = \mathbb{E}_\mu[h_i(\theta)]$. By the L-Lipschitz smoothness of ℓ_i, boundedness $|h_i(\theta)| \le B$, and the definition of F_0, we get for $\mu, \mu' \in \mathcal{P}_2$,

$$\begin{aligned}
F_0(\mu') &\le F_0(\mu) + \frac{1}{n}\sum_{i=1}^n \ell_i'(h_{i,\mu})(h_{i,\mu'} - h_{i,\mu}) + \frac{L}{2n}\sum_{i=1}^n (h_{i,\mu'} - h_{i,\mu})^2 \\
&= F_0(\mu) + \frac{1}{n}\sum_{i=1}^n \ell_i'(h_{i,\mu}) \int h_i(\theta)\mathrm{d}(\mu'-\mu)(\theta) + \frac{L}{2n}\sum_{i=1}^n \left(\int h_i(\theta)\mathrm{d}(\mu'-\mu)(\theta)\right)^2 \\
&\le F_0(\mu) + \int \frac{\delta F_0}{\delta\mu}(\mu)(\theta)\mathrm{d}(\mu'-\mu)(\theta) + \frac{L}{2n}\sum_{i=1}^n \left(\int |h_i(\theta)| \left|\frac{\mathrm{d}(\mu'-\mu)}{\mathrm{d}\theta}(\theta)\right| \mathrm{d}\theta\right)^2 \\
&\le F_0(\mu) + \int \frac{\delta F_0}{\delta\mu}(\mu)(\theta)\mathrm{d}(\mu'-\mu)(\theta) + \frac{B^2 L}{2n}\sum_{i=1}^n \left(\int \left|\frac{\mathrm{d}(\mu'-\mu)}{\mathrm{d}\theta}(\theta)\right| \mathrm{d}\theta\right)^2,
\end{aligned}$$

where we used $\frac{\delta F_0}{\delta\mu}(\mu)(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_i'(h_{i,\mu})h_i(\theta)$.

Since $\hat{\mu}_t$ is the minimizer of the objective (12) with $\mu = \mu^{(t)}$, i.e.,

$$\hat{\mu}^{(t)} = \arg\min_{\mu'} \left\{ \int \frac{\delta F_0}{\delta\mu}(\mu^{(t)})(\theta)\mathrm{d}\mu'(\theta) + \lambda \mathrm{KL}(\mu'\|\nu) \right\},$$

%beginalign*

we see for any $\xi \in \mathcal{P}_2$,

$$\int \frac{\delta F_0}{\delta\mu}(\mu^{(t)})(\theta)\mathrm{d}\hat{\mu}^{(t)}(\theta) \le \int \frac{\delta F_0}{\delta\mu}(\mu^{(t)})(\theta)\mathrm{d}\hat{\mu}^{(t)}(\theta) + \lambda\mathrm{KL}(\hat{\mu}^{(t)}\|\nu) \ \le \int \frac{\delta F_0}{\delta\mu}(\mu^{(t)})(\theta)\mathrm{d}\xi(\theta) + \lambda\mathrm{KL}(\xi\|\nu).$$

Therefore, applying this inequality with $\mu' = (1-\eta\gamma)\mu^{(t)} + \eta\gamma\hat{\mu}^{(t)}$ and $\mu = \mu^{(t)}$, and the convexity of F_0,

$$\begin{aligned}
&F_0(\mu^{(t+1)}) - \eta\gamma\rho \\
&\quad\le F_0((1-\eta\gamma)\mu^{(t)} + \eta\gamma\hat{\mu}^{(t)}) \\
&\quad\le F_0(\mu^{(t)}) + \eta\gamma \int \frac{\delta F_0}{\delta\mu}(\mu^{(t)})(\theta)\mathrm{d}(\hat{\mu}^{(t)} - \mu^{(t)})(\theta) + \frac{\eta^2\gamma^2 B^2 L}{2n}\sum_{i=1}^n \left(\int \left|\frac{\mathrm{d}(\hat{\mu}^{(t)} - \mu^{(t)})}{\mathrm{d}\theta}(\theta)\right| \mathrm{d}\theta\right)^2 \\
&\quad\le F_0(\mu^{(t)}) + \eta\gamma \int \frac{\delta F_0}{\delta\mu}(\mu^{(t)})(\theta)\mathrm{d}(\hat{\mu}^{(t)} - \mu^{(t)})(\theta) + 2\eta^2\gamma^2 B^2 L \\
&\quad\le F_0(\mu^{(t)}) + \eta\gamma \int \frac{\delta F_0}{\delta\mu}(\mu^{(t)})(\theta)\mathrm{d}(\xi - \mu^{(t)})(\theta) + \eta\gamma\lambda\mathrm{KL}(\xi\|\nu) + 2\eta^2\gamma^2 B^2 L \\
&\quad\le F_0(\mu^{(t)}) + \eta\gamma(F_0(\xi) - F_0(\mu^{(t)})) + \eta\gamma\lambda\mathrm{KL}(\xi\|\nu) + 2\eta^2\gamma^2 B^2 L.
\end{aligned}$$

This implies

$$F_0(\mu^{(t+1)}) - F_0(\xi) \leq \eta\gamma\rho + \eta\gamma\lambda \mathrm{KL}(\xi\|\nu) + 2\eta^2\gamma^2 B^2 L + (1-\eta\gamma)(F_0(\mu^{(t)}) - F_0(\xi)).$$

With a slight modification, we get

$$F_0(\mu^{(t+1)}) - F_0(\xi) - \lambda \mathrm{KL}(\xi\|\nu) - \epsilon' \leq (1-\eta\gamma)(F_0(\mu^{(t)}) - F_0(\xi) - \lambda \mathrm{KL}(\xi\|\nu) - \epsilon').$$

where we set $\epsilon' = \rho + 2\eta\gamma B^2 L$. Recursively applying this inequality,

$$F_0(\mu^{(T)}) \leq \epsilon' + F_0(\xi) + \lambda \mathrm{KL}(\xi\|\nu) + (1-\eta\gamma)^T F_0(\mu^{(0)}).$$

Because of arbitrariness of ξ, we conclude

$$F_0(\mu^{(T)}) \leq \epsilon' + (1-\eta\gamma)^T F_0(\mu^{(0)}) + \inf_{\xi \in \mathcal{P}_2} \mathcal{L}(\xi).$$

□

B.2. Convergence analysis of Algorithm 2

We explain that Theorem B.1 can apply to Algorithm 2 as discussed in Remark 6.2. We redefine $\mu^{(t+1)}$ as $\mu^{(t+1)} = (1-\eta\gamma)\mu^{(t)} + \eta\gamma\nu^{(t)}$, where $\nu^{(t)}$ is an empirical distribution $\nu^{(t)} = \frac{1}{m}\sum_{r=1}^m \delta_{\theta_r^{(t)}}$ attained by Langevin Monte Carlo for $\hat{\mu}_t$ with the random initialization in Algorithm 2. Moreover, we set $\overline{\nu}^{(t)} = \mathrm{Law}(\theta_r^{(t)})$. Then, we can evaluate the gap between $\mu^{(t+1)}$ and an exact update $\mu'^{(t+1)} = (1-\eta\gamma)\mu^{(t)} + \eta\gamma\hat{\mu}^{(t)}$ as follows. Suppose $\|h_i\|_\infty \leq B$ and ℓ_i is C-Lipschitz continuous. Then, we get

$$\begin{aligned}
|F_0(\mu'^{(t+1)}) - F_0(\mu^{(t+1)})| &\leq \frac{C}{n}\sum_{i=1}^n \left|\mathbb{E}_{\hat{\mu}'^{(t+1)}}[h_i] - \mathbb{E}_{\mu^{(t+1)}}[h_i]\right| \\
&= \frac{\eta\gamma C}{n}\sum_{i=1}^n \left|\mathbb{E}_{\hat{\mu}^{(t)}}[h_i] - \mathbb{E}_{\nu^{(t)}}[h_i]\right| \\
&\leq \frac{\eta\gamma C}{n}\sum_{i=1}^n \left|\mathbb{E}_{\hat{\mu}^{(t)}}[h_i] - \mathbb{E}_{\overline{\nu}^{(t)}}[h_i]\right| + \frac{\eta\gamma C}{n}\sum_{i=1}^n |\mathbb{E}_{\overline{\nu}^{(t)}}[h_i] - \mathbb{E}_{\nu^{(t)}}[h_i]| \\
&\leq \eta\gamma BC \left\| \frac{\mathrm{d}\hat{\mu}^{(t)}}{\mathrm{d}\theta} - \frac{\mathrm{d}\overline{\nu}^{(t)}}{\mathrm{d}\theta} \right\|_{L_1(\mathrm{d}\theta)} + \frac{\eta\gamma C}{n}\sum_{i=1}^n |\mathbb{E}_{\overline{\nu}^{(t)}}[h_i] - \mathbb{E}_{\nu^{(t)}}[h_i]| \\
&\leq \eta\gamma BC\sqrt{2\mathrm{KL}(\overline{\nu}^{(t)}\|\hat{\mu}^{(t)})} + \frac{\eta\gamma C}{n}\sum_{i=1}^n |\mathbb{E}_{\overline{\nu}^{(t)}}[h_i] - \mathbb{E}_{\nu^{(t)}}[h_i]|,
\end{aligned} \tag{25}$$

where we applied Pinsker's inequality for the last inequality.

Therefore, we can estimate the particle and iteration complexities to satisfy the required precision (24) by applying the convergence rate of Langevin Monte Carlo (Vempala & Wibisono, 2019) together with a standard concentration inequality as done in Nitanda et al. (2021); Oko et al. (2022).

Definition B.2 (Log-Sobolev inequality). Let $\mathrm{d}\mu = p(\theta)\mathrm{d}\theta$ be a probability distribution with a positive smooth density $p > 0$ on $\mathbb{R}^p$. We say that μ satisfies the log-Sobolev inequality with constant $\alpha > 0$ if for any smooth function $f : \mathbb{R}^p \to \mathbb{R}$,

$$\mathbb{E}_\mu[f^2 \log f^2] - \mathbb{E}_\mu[f^2]\log \mathbb{E}_\mu[f^2] \leq \frac{2}{\alpha}\mathbb{E}_\mu[\|\nabla f\|_2^2].$$

This inequality is analogous to strong convexity in optimization: let $\mathrm{d}\nu = q(\theta)\mathrm{d}\mu$ be a probability distribution on $\mathbb{R}^p$ such that q is smooth and positive. Then, if μ satisfies the log-Sobolev inequality with α, it follows that

$$\mathrm{KL}(\nu||\mu) \leq \frac{1}{2\alpha}\mathbb{E}_\nu[\|\nabla_\theta \log q\|_2^2].$$

The above relation is directly obtained by setting $f = \sqrt{q}$ in the definition of log-Sobolev inequality. Note that the right hand side is nothing else but the squared norm of functional gradient of $\mathrm{KL}(\nu\|\mu)$ with respect to a transport map for ν.

It is well-known that strong log-concave densities satisfy the LSI with a dimension-free constant (up to the spectral norm of the covariance).

Example B.3 ((Bakry & Émery, 1985)). Let $q \propto \exp(-f)$ be a probability density, where $f : \mathbb{R}^p \to \mathbb{R}$ is a smooth function. If there exists $c > 0$ such that $\nabla^2 f \succeq cI_p$, then $q(\theta)\mathrm{d}\theta$ satisfies Log-Sobolev inequality with constant c.

In addition, the LSI is preserved under bounded perturbation, as originally shown in (Holley & Stroock, 1987). We also provide a proof for completeness.

Lemma B.4 ((Holley & Stroock, 1987)). *Let $q(\theta)\mathrm{d}\theta$ be a probability distribution on $\mathbb{R}^p$ satisfying the log-Sobolev inequality with a constant α. For a bounded function $B : \mathbb{R}^p \to \mathbb{R}$, we define a probability distribution $q_B(\theta)\mathrm{d}\theta$ as follows:*

$$q_B(\theta)\mathrm{d}\theta = \frac{\exp(B(\theta))q(\theta)}{\mathbb{E}_q[\exp(B(\theta))]}\mathrm{d}\theta.$$

Then, $q_B\mathrm{d}\theta$ satisfies the log-Sobolev inequality with a constant $\alpha/\exp(4\|B\|_\infty)$.

Theorem B.5 ((Vempala & Wibisono, 2019)). *Consider a probability density $q(\theta) \propto \exp(-f(\theta))$ satisfying the log-Sobolev inequality with constant α, and assume f is smooth and ∇f is L-Lipschitz, i.e., $\|\nabla_\theta f(\theta) - \nabla_\theta f(\theta')\|_2 \le L\|\theta - \theta'\|_2$. If we run the unadjusted Langevin algorithm with learning rate $0 < \eta' \le \frac{\alpha}{4L^2}$ and let $q^{(s)}(\theta)\mathrm{d}\theta$ be a probability distribution that $\theta^{(s)}$ follows, then we have,*

$$\mathrm{KL}(q^{(s)}\|q) \le \exp(-\alpha\eta' s)\mathrm{KL}(q^{(1)}\|q) + 8\alpha^{-1}\eta' dL^2.$$

Lemma B.6 (Hoeffding's inequality). *Let $Z, Z_1, \ldots, Z_m$ be i.i.d. random variables taking values in $[-a, a]$ for $a > 0$. Then, for any $\rho > 0$, we get*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{r=1}^{m} Z_r - \mathbb{E}[Z]\right| > \rho\right] \le 2\exp\left(-\frac{\rho^2 m}{2a^2}\right).$$

We evaluate the inequality (25) by combining the above auxiliary results. By Lemma B.4 with Example B.3, we see at least $\hat{\mu}^{(t)}$ satisfies LSI with the constant $\alpha = \frac{2\lambda'}{\lambda}\exp\left(-\frac{4BC}{\lambda}\right)$. Furthermore, under the Lipschitz continuity of ∇h_i where $\|\nabla h_i(\theta') - \nabla h_i(\theta)\|_2 \le D\|\theta' - \theta\|_2$ for any $\theta, \theta' \in \mathbb{R}^p$, we see $\nabla \log \frac{\mathrm{d}\hat{\mu}^{(t)}}{\mathrm{d}\theta}$ is Lipschitz continuous with the constant $L' = \frac{1}{\lambda}(CD + 2\lambda')$. We run Langevin Monte Carlo for $\hat{\mu}^{(t)}$ with the initialization $\overline{\theta}_r^{(0)} \sim \mu^{(0)}$. Then, if $S \ge \frac{1}{\alpha\eta'}\log\frac{\alpha\mathrm{KL}(\mu^{(0)}\|\hat{\mu}^{(t)})}{8\eta' dL'^2}$, we have by Theorem B.5,

$$\mathrm{KL}(q^{(S)}\|\hat{\mu}^{(t)}) \le 16\alpha^{-1}\eta' dL'^2,$$

where $q^{(S)}$ is the distribution of $\overline{\theta}_r^{(S)}$. Moreover, by Lemma B.6, we see with probability at least $1 - \delta$, for any $i \in \{1, 2, \ldots, n\}$,

$$|\mathbb{E}_{\overline{\nu}^{(t)}}[h_i] - \mathbb{E}_{\nu^{(t)}}[h_i]| \le B\sqrt{\frac{2}{m}\log\frac{2n}{\delta}}.$$

Noting $\theta_r^{(t)} = \overline{\theta}_r^{(S)}$ (i.e., $\overline{\nu}^{(t)} = q^{(S)}$), we get with probability at least $1 - \delta$,

$$\begin{aligned}|F_0(\mu'^{(t+1)}) - F_0(\mu^{(t+1)})| &\le \eta\gamma BC\sqrt{2\mathrm{KL}(\overline{\nu}^{(t)}\|\hat{\mu}^{(t)})} + \frac{\eta\gamma C}{n}\sum_{i=1}^{n}|\mathbb{E}_{\overline{\nu}^{(t)}}[h_i] - \mathbb{E}_{\nu^{(t)}}[h_i]| \\ &\le \eta\gamma\left(4BCL'\sqrt{\frac{2\eta' d}{\alpha}} + BC\sqrt{\frac{2}{m}\log\frac{2n}{\delta}}\right).\end{aligned}$$

Note this is a high probability bound per outer iteration t. Therefore, we take a union bound with respect to $t \in \{0, 1, \ldots, T-1\}$ to obtain high probability bounds that hold all t. That is, with probability at least $1 - \delta$, for any

$t \in \{0, 1, \ldots, T-1\}$,

$$|F_0(\mu'^{(t+1)}) - F_0(\mu^{(t+1)})| \leq \eta\gamma \left(4BCL'\sqrt{\frac{2\eta' d}{\alpha}} + BC\sqrt{\frac{2}{m}\log\frac{2nT}{\delta}} \right).$$

To satisfy the condition (24) at all iterations t under the above event, it is sufficient take η' and m as follows:

$$\eta' \leq \frac{\alpha\rho^2}{128dB^2C^2L'^2},\ m \geq \frac{8B^2C^2}{\rho^2}\log\frac{2nT}{\delta}.$$

Taking everything into account, we get the following convergence guarantee for Algorithm 2.

Theorem B.7. *Suppose Assumption 2.1 holds, ℓ_i is C-Lipschitz continuous and L-Lipschitz smooth, $\ell_i \geq 0$, and $\|h\|_\infty \leq B$, and ∇h_i is D-Lipschitz continuous regarding $\|\cdot\|_2$. Suppose the underlying Langevin Monte Carlo is run with the initialization $\overline{\theta}_r^{(0)} \sim \mu^{(0)}$. Given $\delta \in (0,1)$, hyperparameters η', S, m of Algorithm 2, are set as follows:*

$$\eta' \leq \min\left\{\frac{\alpha}{4L'^2}, \frac{\alpha\rho^2}{128dB^2C^2L'^2}\right\},\ S \geq \frac{1}{\alpha\eta'}\log\frac{\alpha\mathrm{KL}(\mu^{(0)}\|\hat{\mu}^{(t)})}{8\eta' dL'^2},\ m \geq \frac{8B^2C^2}{\rho^2}\log\frac{2nT}{\delta},$$

where α is the LSI-constant of $\hat{\mu}^{(t)}$ and $L' = \frac{1}{\lambda}(CD + 2\lambda')$. Then, for $\{\mu^{(t)}\}_{t=0}^T \subset \mathcal{P}_2$ generate by Algorithm 2, we get with probability at least $1-\delta$,

$$F_0(\mu^{(T)}) \leq \epsilon' + (1-\eta\gamma)^T F_0(\mu^{(0)}) + \inf_{\xi\in\mathcal{P}_2}\mathcal{L}(\xi),$$

where $\epsilon' = \rho + 2\eta\gamma B^2 L$.

We here derive particle and iteration complexities to achieve a given ϵ-accurate solution in the following sense: $F_0(\mu^{(T)}) \leq \inf_{\xi\in\mathcal{P}_2}\mathcal{L}(\xi) + \epsilon$. To this end, we set $\rho = \Theta(\epsilon),\ \eta\gamma = \Theta(\epsilon)$, and $T = \Theta\left(\epsilon^{-1}\log\frac{F_0(\mu^{(0)})}{\epsilon}\right) = \tilde{\Theta}(\epsilon^{-1})$. The notation $\tilde{\Theta}$ hides the logarithmic term for simplicity. For the step size η', it is sufficient to consider the case of $\eta' = \Theta\left(\alpha\lambda^2\epsilon^2\right)$. Then, $S = \Theta\left(\alpha^{-2}\lambda^{-2}\epsilon^{-2}\log\frac{\mathrm{KL}(\mu^{(0)}\|\hat{\mu}^{(t)})}{8\epsilon^2 dL^2}\right) = \tilde{\Theta}(\alpha^{-2}\lambda^{-2}\epsilon^{-2})$ and $m = \Theta\left(\epsilon^{-2}\log\frac{2nT}{\delta}\right) = \tilde{\Theta}(\epsilon^{-2})$. Therefore, particle and iteration complexities are as follows, respectively:

$$m = \tilde{\Theta}(\epsilon^{-2}),\ ST = \tilde{\Theta}(\alpha^{-2}\lambda^{-2}\epsilon^{-3}).$$