Convex Analysis of the Mean Field Langevin Dynamics

Atsushi Nitanda[†]
nitanda@ai.kyutech.ac.jp

Denny Wu[‡]
dennywu@cs.toronto.edu

Taiji Suzuki[⋆]
taiji@mist.i.u-tokyo.ac.jp

† Kyushu Institute of Technology and RIKEN Center for Advanced Intelligence Project
‡ University of Toronto and Vector Institute for Artificial Intelligence
⋆ University of Tokyo and RIKEN Center for Advanced Intelligence Project

Abstract

As an example of the nonlinear Fokker-Planck equation, the *mean field Langevin dynamics* recently attracts attention due to its connection to (noisy) gradient descent on infinitely wide neural networks in the mean field regime, and hence the convergence property of the dynamics is of great theoretical interest. In this work, we give a concise and self-contained convergence rate analysis of the mean field Langevin dynamics with respect to the (regularized) objective function in both continuous and discrete time settings. The key ingredient of our proof is a *proximal Gibbs distribution* p_q associated with the dynamics, which, in combination with techniques in Vempala and Wibisono (2019), allows us to develop a simple convergence theory parallel to classical results in convex optimization. Furthermore, we reveal that p_q connects to the *duality gap* in the empirical risk minimization setting, which enables efficient empirical evaluation of the algorithm convergence.

1 INTRODUCTION

Consider a neural network with M trainable neurons parameterized as

$$h_\Theta(x) \overset{\text{def}}{=} \frac{1}{M} \sum_{r=1}^{M} h_{\theta_r}(x), \qquad (1)$$

where each neuron h_{θ_r} contains trainable parameters (weights) θ_r and some nonlinear transformation (e.g.,

$h_{\theta_r}(x) = \sigma(\langle \theta_r, x \rangle)$ where σ is the nonlinear activation function), x is a data example, and $\Theta = \{\theta_r\}_{r=1}^{M}$. Under suitable conditions, as $M \to \infty$ we obtain the *mean field limit*: $h_q(x) := \mathbb{E}_q[h_\theta(x)]$, where $q(\theta)\mathrm{d}\theta$ represents the probability distribution of the weights; we refer to h_q simply as a mean field model (neural network). In this limit, training can be formulated as an optimization problem over the space of probability measures. An advantage of the mean field regime, in contrast to alternative settings such as the neural tangent kernel regime (Jacot et al., 2018), is the presence of (nonlinear) *feature learning* (Suzuki, 2019; Ghorbani et al., 2019). However, developing an optimization theory for mean field neural networks is also more challenging.

Optimization analysis of mean field neural networks usually utilizes the convexity of the objective function in the space of probability measures. Nitanda and Suzuki (2017); Chizat and Bach (2018); Mei et al. (2018) established global convergence of gradient descent (flow) on two-layer neural networks in the mean field regime under appropriate conditions. Subsequent works proved convergence rates under additional structural assumptions (Javanmard et al., 2019; Chizat, 2021b). One noticeable algorithmic modification is the addition of Gaussian noise to the gradient, which leads to the *noisy* gradient descent algorithm; this modification gives rise to an entropy regularization term in the objective, and allows for global convergence analysis under less restrictive settings (Rotskoff and Vanden-Eijnden, 2018; Mei et al., 2019). The corresponding stochastic dynamics is often referred to as the *mean field Langevin dynamics*.

Recent works have studied the convergence rate of the mean field Langevin dynamics, including its underdamped (kinetic) version. However, most existing analyses either require sufficiently strong regularization (Hu et al., 2019; Jabir et al., 2019), or build upon involved mathematical tools (Kazeykina et al., 2020; Guillin et al., 2021). Our goal is to provide a simpler convergence proof that covers general and more

practical machine learning settings, with a focus on neural network optimization in the mean field regime. Motivated by an observation in Nitanda et al. (2021) that the *log-Sobolev inequality* can simplify the global convergence analysis of two-layer mean field neural network, we study the optimization efficiency of the mean field Langevin dynamics in the context of KL-regularized empirical/expected risk minimization, and present a new convergence rate analysis by translating the finite-dimensional convex optimization theory to the optimization in the space of measures.
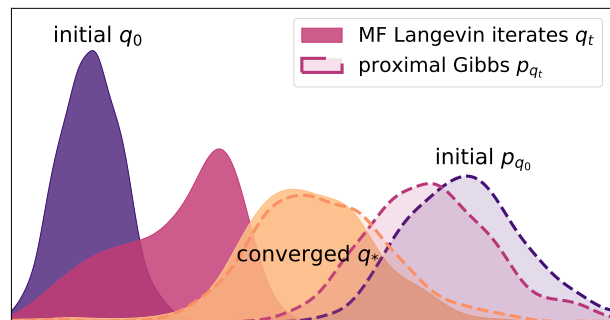

Figure 1: 1D visualization of mean field two-layer neural network (tanh) optimized by noisy gradient descent ($\lambda = \lambda' = 10^{-2}$). Observe that both the parameter distribution q and the corresponding proximal Gibbs distribution p_q converge to the optimal q_*. Moreover, q_t and p_{q_t} approach q_* from opposite directions, as predicted by Proposition 1.

1.1 Contributions

In this work, we give a simple and self-contained convergence rate analysis of the mean field Langevin dynamics in both continuous and discrete time settings. The key ingredient of our proof is the introduction of a *proximal Gibbs distribution* p_q (see Figure 1), which relates to the optimization gap and allows us to directly apply standard convex optimization techniques. In particular,

- By analyzing the proximal Gibbs distribution p_q, we establish linear convergence in continuous time with respect to the KL-regularized objective. This convergence result holds for *any* regularization parameters, in contrast to existing analyses (e.g., Hu et al. (2019)) that require strong regularization.

- We also provide global convergence rate analysis for the discrete-time update. This is achieved by extending the classical "one-step interpolation" argument in the analysis of Langevin dynamics (e.g., see Vempala and Wibisono (2019) for the KL case) to our nonlinear Fokker-Planck setting.

- Finally, we present an interpretation of the proximal Gibbs distribution via the primal-dual formulation of empirical risk minimization problems, and reveal that p_q exactly fills the duality gap. This interpretation

leads to alternative ways to evaluate the convergence of the algorithm.

1.2 Related Literature

Convergence of the Langevin algorithm. The Langevin dynamics can be interpreted as the (Wasserstein) gradient flow of KL divergence with respect to the target distribution $p \propto \exp(-f)$ (Jordan et al., 1998); in other words, the Langevin dynamics solves the following optimization problem,

$$\min_{q:\text{density}} \left\{ \mathbb{E}_q[f] + \mathbb{E}_q[\log(q)] \right\}. \qquad (2)$$

We refer readers to Dalalyan (2017); Wibisono (2018) for additional discussions on the connection between sampling and optimization.

It is well-known that when the target distribution satisfies certain isoperimetry conditions such as the log-Sobolev inequality, then the (continuous-time) Langevin dynamics converges exponentially (Bakry et al., 2013). On the other hand, the time-discretized dynamics, or the Langevin algorithm, admits a biased invariant distribution depending on the step size and the numerical scheme (Milstein and Tretyakov, 2013; Li et al., 2019), and sublinear convergence rate has been established under different target assumptions (Vempala and Wibisono, 2019; Erdogdu and Hosseinzadeh, 2020) and distance metrics (Dalalyan, 2014; Durmus and Moulines, 2017; Erdogdu et al., 2021).

Mean field regime and nonlinear Fokker-Planck. Analysis of neural networks in the mean field regime typically describes the optimization dynamics as a partial differential equation (PDE) of the parameter distribution, from which convergence to the global optimal solution may be shown (Chizat and Bach, 2018; Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018; Sirignano and Spiliopoulos, 2020). Quantitative convergence rate usually requires additional conditions, such as structural assumptions on the learning problem (Javanmard et al., 2019; Chizat, 2021b; Akiyama and Suzuki, 2021), or modification of the dynamics (Rotskoff et al., 2019; Wei et al., 2019). Noticeably, Mei et al. (2018); Hu et al. (2019); Chen et al. (2020a) considered the optimization of the *KL-regularized objective*, which leads to the mean field Langevin dynamics.

Note that the optimization of mean field neural networks (e.g., Eq. (13)) falls beyond the scope of Langevin dynamics whose density function follows *linear* Fokker-Planck equation and solves (2), due to nonlinear loss function. Instead, the density function of parameters follows a *nonlinear* Fokker-Planck equation (Eq. (8)), the convergence rate of which is more difficult to establish. Hu et al. (2019); Jabir et al. (2019) obtained

convergence rate of the mean field Langevin dynamics under sufficiently strong regularization. Exponential convergence of related dynamics (e.g., the underdamped variant and other McKean-Vlasov equations) in continuous time have been shown under various settings (Monmarché, 2017; Guillin et al., 2019; Kazeykina et al., 2020; Guillin et al., 2021), based on hypocoercivity (Villani, 2009) or coupling techniques (Eberle et al., 2019). Bou-Rabee and Schuh (2020); Bou-Rabee and Eberle (2021) studied the discrete time convergence of Hamiltonian Monte Carlo with interaction potential.

Our work builds upon Nitanda et al. (2021), which employs the Langevin algorithm to solve the inner loop of a *dual averaging* method in the space of measures. Importantly, under a uniform log-Sobolev inequality on the "linearized" objective, (sublinear) global convergence rate in minimizing an entropy-regularized nonlinear functional can be proved by adapting finite-dimensional convex optimization theory. Based on similar ideas combined with the primal-dual formulation of empirical risk minimization, a *stochastic dual coordinate ascent* alternative has been developed in Oko et al. (2022) to achieve linear convergence in discrete time. A few recent works also considered the adaptation of classical convex optimization algorithms into the space of measures, such as the Mirror descent method (Ying, 2020), the Frank-Wolfe method (Kent et al., 2021), and the Bregman proximal gradient method (Chizat, 2021a).

Concurrent and independent to our work, Chizat (2022) analyzed the mean field Langevin dynamics also using properties of the proximal Gibbs distribution p_q; while both works build upon the same observation, Chizat (2022) focused on the continuous-time convergence rate and studied the annealed dynamics, whereas we establish discrete-time guarantees for the noisy gradient descent algorithm and present a primal-dual viewpoint.

1.3 Notations

$\|\cdot\|_2$ denotes the Euclidean norm. Given a density function $q(\theta)$ on $\mathbb{R}^d$, we write the expectation w.r.t. $q(\theta)d\theta$ as $\mathbb{E}_{\theta \sim q}[\cdot]$ or simply $\mathbb{E}_q[\cdot]$, $\mathbb{E}_\theta[\cdot]$ when the random variable and distribution are obvious from the context; e.g. for a function $f : \mathbb{R}^d \to \mathbb{R}$, we write $\mathbb{E}_q[f] = \int f(\theta)q(\theta)d\theta$ when f is integrable. KL stands for the Kullback-Leibler divergence: $\mathrm{KL}(q\|q') \stackrel{\text{def}}{=} \int q(\theta) \log\left(\frac{q(\theta)}{q'(\theta)}\right) d\theta$. Let $\mathcal{P}$ be the space of probability density functions with respect to $d\theta$ on $\mathbb{R}^d$ such that the entropy and second moment are well-defined.

2 PRELIMINARIES

The mean field Langevin dynamics is the target of our convergence analysis. In this section, we introduce this dynamics as well as the associated optimization problem. We also outline one major application of our convergence analysis, which is the optimization of mean field neural networks.

2.1 Problem Setup

Let $F : \mathcal{P} \to \mathbb{R}$ be a differentiable convex functional. That is, we suppose there is a functional $\frac{\delta F}{\delta q} : \mathcal{P} \times \mathbb{R}^d \ni (q, \theta) \mapsto \frac{\delta F}{\delta q}(q)(\theta) \in \mathbb{R}$ such that for any $q, q' \in \mathcal{P}$,

$$\frac{dF(q + \epsilon(q' - q))}{d\epsilon}\bigg|_{\epsilon=0} = \int \frac{\delta F}{\delta q}(q)(\theta)(q' - q)(\theta)d\theta$$

and F satisfies the convexity condition:

$$F(q') \geq F(q) + \int \frac{\delta F}{\delta q}(q)(\theta)(q' - q)(\theta)d\theta. \quad (3)$$

We consider the minimization of an entropy regularized nonlinear functional:

$$\min_{q \in \mathcal{P}} \left\{ \mathcal{L}(q) \stackrel{\text{def}}{=} F(q) + \lambda \mathbb{E}_q[\log q] \right\}. \quad (4)$$

For $q \in \mathcal{P}$, we next define an associated Gibbs distribution which plays a key role in our analysis: the *proximal Gibbs distribution* around q in $\mathcal{P}$ (see Proposition 1).

Definition 1 (Proximal Gibbs distribution). *We define $p_q(\theta)$ to be the Gibbs distribution with potential function $-\lambda^{-1}\delta F(q)/\delta q$, i.e.,*

$$p_q(\theta) = \frac{\exp\left(-\frac{1}{\lambda}\frac{\delta F(q)}{\delta q}(\theta)\right)}{Z(q)}, \quad (5)$$

where $Z(q)$ is the normalization constant.

In this study, we provide basic convergence results for the general form of the problem (4) under the following assumptions, which will be specialized and verified for mean field neural network introduced in the sequel.

Assumption 1. *For a functional $F : \mathcal{P} \to \mathbb{R}$ and any $q \in \mathcal{P}$, assume the functional derivative $\frac{\delta F}{\delta q}(q)(\theta)$ exists and is smooth in θ. Moreover, assume $|Z(q)| < \infty$, $\frac{\delta F}{\delta q}(q)(\theta) = O(1 + \|\theta\|_2^2)$ uniformly over $\mathcal{P}$, and F is a convex functional, that is, (3) holds for any $q, q' \in \mathcal{P}$.*

Assumption 2 (Log-Sobolev inequality). *Suppose there exists a constant $\alpha > 0$ such that for any $q \in \mathcal{P}$, the probability distribution $p_q(\theta)d\theta$ satisfies log-Sobolev inequality with constant α, that is, for any smooth function $g : \mathbb{R}^d \to \mathbb{R}$, we have*

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

Under Assumption 2, by setting $g = \sqrt{q/p_q}$, we get

$$\mathrm{KL}(q\|p_q) \leq \frac{1}{2\alpha}\mathbb{E}_q\left[\left\|\nabla \log \frac{q}{p_q}\right\|_2^2\right]. \quad (6)$$

2.2 Optimization Dynamics

To solve the problem (4), we consider the following (continuous-time) *mean field Langevin dynamics*:

$$d\theta_t = -\nabla \frac{\delta F}{\delta q}(q_t)(\theta_t)dt + \sqrt{2\lambda}dW_t, \qquad (7)$$

where $\theta_t \sim q_t(\theta)d\theta$ and $\{W_t\}_{t\geq 0}$ is the Brownian motion in $\mathbb{R}^d$ with $W_0 = 0$. Here, the gradient ∇ of the functional derivative $\frac{\delta F(q)}{\delta q}(\theta)$ is applied with respect to θ. It is known that the distribution of θ_t following the dynamics (7) solves nonlinear Fokker–Planck equation:

$$\frac{\partial q_t}{\partial t} = \nabla \cdot \left(q_t \nabla \frac{\delta F}{\delta q}(q_t) \right) + \lambda \Delta q_t. \qquad (8)$$

We here reformulate the equation (8) as follows:

$$\frac{\partial q_t}{\partial t} = \lambda \nabla \cdot \left(q_t \nabla \log \exp \left(\frac{1}{\lambda} \frac{\delta F}{\delta q}(q_t) \right) + q_t \nabla \log q_t \right)$$

$$= \lambda \nabla \cdot \left(q_t \nabla \log \frac{q_t}{p_{q_t}} \right). \qquad (9)$$

As we will see later, this formulation involving p_{q_t} will be more useful in the convergence analysis.

Moreover, we also consider the standard discretization of the above dynamics with step size $\eta > 0$:

$$\theta^{(k+1)} = \theta^{(k)} - \eta \nabla \frac{\delta F}{\delta q}(q^{(k)})(\theta^{(k)}) + \sqrt{2\lambda\eta}\xi^{(k)}, \quad (10)$$

where $\theta^{(k)}$ is a random variable following $q^{(k)}(\theta)d\theta$ and $\xi^{(k)} \sim \mathcal{N}(0, I_d)$. We present convergence rate analysis for both the continuous- and discrete-time algorithm.

2.3 Mean Field Neural Networks

The main application of our theory is to provide quantitative convergence guarantees for risk minimization problems with the mean field neural network. Before formally defining the mean field limit, we first introduce the finite dimensional counterpart.

Finite dimensional case. Let $\mathcal{X}$ be the data space and $h_\theta : \mathcal{X} \to \mathbb{R}$ be a component of neural network, which corresponds to a single neuron with trainable parameter $\theta \in \mathbb{R}^d$. Then, an M-neuron network h_Θ (where $\Theta = \{\theta_r\}_{r=1}^M$) is defined as the average of these components as in (1). Let $\ell(z, y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a smooth and convex loss function in z, such as the logistic loss and squared loss, and ρ be the empirical or true data distribution. Then, the regularized risk of h_Θ is defined as

$$\mathbb{E}_{(X,Y)\sim\rho}[\ell(h_\Theta(X), Y)] + \frac{\lambda'}{M} \sum_{r=1}^M r(\theta_r), \qquad (11)$$

where $r(\theta)$ is a convex regularizer such as ℓ_2-penalty $r(\theta) = \|\theta\|_2^2$, and $\lambda' > 0$ is regularization strength. Note that this formulation includes both empirical and expected risk minimization problems. To optimize the objective (11), we perform the gradient descent update with step size $\eta M > 0$:

$$g_r^{(k)} = \mathbb{E}[\partial_z \ell(h_{\Theta^{(k)}}(X), Y) \partial_{\theta_r} h_{\theta_r^{(k)}}(X)] + \lambda' \partial_{\theta_r} r(\theta_r^{(k)}),$$

$$\theta_r^{(k+1)} = \theta_r^{(k)} - \eta g_r^{(k)}. \qquad (12)$$

Mean field limit. We now take the limit $M \to \infty$ and suppose θ_r follows a probability distribution $q(\theta)d\theta$. Then, $h_\Theta(x)$ converges to the mean field limit $h_q(x) = \mathbb{E}_{\theta\sim q}[h_\theta(x)]$ in which the density function q is recognized as the parameter, and the objective (11) converges to the following convex functional,

$$F(q) = \mathbb{E}_{(X,Y)}[\ell(h_q(X), Y)] + \lambda' \mathbb{E}_{\theta\sim q}[r(\theta)]. \quad (13)$$

In this case, the functional derivative $\frac{\delta F(q)}{\delta q}$ is

$$\frac{\delta F(q)}{\delta q}(\theta) = \mathbb{E}_{(X,Y)}[\partial_z \ell(h_q(X), Y)h_\theta(X)] + \lambda' r(\theta).$$

By noticing that $\nabla \frac{\delta F(q)}{\delta q}(\theta)$ is the mean field limit $(M \to \infty)$ of the step $g_r^{(k)}$ for the finite-dimensional model, we see that the discrete dynamics (10) is the noisy variant of the mean field limit of gradient descent (12). This motivates us to study mean field Langevin dynamics (7) and its discretization (10).

Verification of assumptions. Under smoothness and boundedness conditions on h_θ, and also smooth convex loss function ℓ with ℓ_2-penalty term $r(\theta) = \|\theta\|_2^2$, one can easily verify that objective of mean field neural network satisfies Assumption 1. The convexity of F immediately follows by taking the expectation $\mathbb{E}_{(X,Y)}$ of the following inequality:

$$\ell(h_q(X), Y) + \int \partial_z \ell(h_q(X), Y)h_\theta(X)(q' - q)(\theta)d\theta$$

$$= \ell(h_q(X), Y) + \partial_z \ell(h_q(X), Y)(h_{q'}(X) - h_q(X))$$

$$\leq \ell(h_{q'}(X), Y).$$

Moreover, with $r(\theta) = \|\theta\|_2^2$, the uniform log-Sobolev inequality (Assumption 2) with constant $\alpha = \frac{2\lambda'}{\lambda \exp(O(\lambda^{-1}))}$ can be verified via a standard application of the LSI perturbation lemma (Holley and Stroock, 1987) (see Appendix A. 1 for details), since p_q is proportional to the Gibbs distribution specified as:

$$\exp \left(-\frac{1}{\lambda} \mathbb{E}_{(X,Y)}[\partial_z \ell(h_q(X), Y)h_\theta(X)] - \frac{\lambda'}{\lambda} r(\theta) \right). \qquad (14)$$

3 CONVERGENCE ANALYSIS

In this section, we present the convergence rate analysis of mean field Langevin dynamics in both continuous- and discrete-time settings. We remark that our proof strategy can be seen as a combination and extension of (i) convex analysis parallel to the finite-dimensional optimization setting, and (ii) convergence analysis for the linear Fokker-Planck equation (e.g., Vempala and Wibisono (2019)). As we will see, the proximal Gibbs distribution p_q plays an important role in connecting these different techniques.

3.1 Basic Properties and Convexity

We first present a basic but important result that characterizes the role of p_q in the convergence analysis. Note that the functional derivative of the negative entropy $\mathbb{E}_q[\log q]$ in the density function q is $\log q$[1]; therefore, the optimality condition of the problem (4) is $\frac{\delta \mathcal{L}}{\delta q}(q) = \frac{\delta F}{\delta q}(q) + \lambda \log q = 0$. This is to say, the optimal probability density function q_* satisfies

$$q_*(\theta) = p_{q_*}(\theta) \propto \exp\left(-\frac{1}{\lambda}\frac{\delta F(q_*)}{\delta q}(\theta)\right). \quad (15)$$

This fact was shown in Hu et al. (2019). Hence, we may interpret the divergence between probability density functions q and p_q as an *optimization gap*. Indeed, this intuition is confirmed by the following proposition, which can be established using standard convex analysis. In particular, the proof relies on the fact that the negative entropy acts as a strongly convex function with respect to KL-divergence.

Proposition 1. *Under Assumption 1, we have the three following statements,*

1. In the sense of functional on the space of probability density functions, the following equality holds.

$$\frac{\delta \mathcal{L}}{\delta q}(q) = \lambda \frac{\delta}{\delta q'}\mathrm{KL}(q'\|p_q)|_{q'=q} = \lambda \log \frac{q}{p_q}.$$

In other words, for any $g = p - p'$ $(p, p' \in \mathcal{P})$,

$$\int \frac{\delta \mathcal{L}}{\delta q}(q)(\theta)g(\theta)\mathrm{d}\theta = \int \lambda \log\left(\frac{q}{p_q}(\theta)\right)g(\theta)\mathrm{d}\theta.$$

2. For any probability distributions $q, q' \in \mathcal{P}$, we have

$$\mathcal{L}(q) + \int \frac{\delta \mathcal{L}}{\delta q}(q)(\theta)(q'-q)(\theta)\mathrm{d}\theta + \lambda \mathrm{KL}(q'\|q) \le \mathcal{L}(q'). \quad (16)$$

Moreover, p_q associated with $q \in \mathcal{P}$ is a minimizer of the left hand side of this inequality in $q' \in \mathcal{P}$.

[1]We ignore a constant in the functional derivative because this derivative is applied to function $g : \mathbb{R}^d \to \mathbb{R}$ that satisfies $\int g(\theta)\mathrm{d}\theta = 0$ or is differentiated by the gradient ∇.

3. Let q_* be an optimal solution of (4). Then, for any $q \in \mathcal{P}$, we get

$$\lambda \mathrm{KL}(q\|p_q) \ge \mathcal{L}(q) - \mathcal{L}(q_*) \ge \lambda \mathrm{KL}(q\|q_*).$$

Proof. (i) We start with the first statement.

$$
\begin{aligned}
\frac{\delta \mathcal{L}}{\delta q}(q) &= \frac{\delta F}{\delta q}(q) + \lambda \log q \\
&= -\lambda \log \exp\left(-\frac{1}{\lambda}\frac{\delta F}{\delta q}(q)\right) + \lambda \log q \\
&= \lambda(\log q - \log p_q) - \lambda \log Z(p_q),
\end{aligned}
$$

where $Z(p_q)$ is a normalization constant of p_q. Moreover, we can see $\frac{\delta}{\delta q'}\mathrm{KL}(q'\|p_q)|_{q'=q} = \log q - \log p_q$.

(ii) From direct computation, for any $q, q' \in \mathcal{P}$,

$$
\begin{aligned}
\mathbb{E}_{q'}[\log(q')] = &\mathbb{E}_q[\log(q)] + \mathrm{KL}(q'\|q) \\
&+ \int \frac{\delta}{\delta q}\mathbb{E}_q[\log(q)](\theta)(q' - q)(\theta)\mathrm{d}\theta. \quad (17)
\end{aligned}
$$

By the convexity of F and (17), we get

$$
\begin{aligned}
\mathcal{L}(q') &= F(q') + \lambda \mathbb{E}_{q'}[\log q'] \\
&\ge F(q) + \int \frac{\delta F}{\delta q}(q)(\theta)(q' - q)(\theta)\mathrm{d}\theta + \lambda \mathbb{E}_{q'}[\log q'] \\
&= \mathcal{L}(q) + \int \frac{\delta \mathcal{L}}{\delta q}(q)(\theta)(q' - q)(\theta)\mathrm{d}\theta + \lambda \mathrm{KL}(q'\|q).
\end{aligned}
$$

In addition, by taking the functional derivative of the left hand side of (16) in q', we obtain the optimality condition of this functional as follows:

$$0 = \frac{\delta \mathcal{L}}{\delta q} + \lambda(\log q' - \log q) = -\lambda \log p_q + \lambda \log q'.$$

Therefore, $q' = p_q$ is a minimizer of the left hand side of (16) as desired.

(iii) For the last statement, observe that minimizing both sides of (16) over $q' \in \mathcal{P}$ yields

$$
\begin{aligned}
\mathcal{L}(q_*) &\ge \mathcal{L}(q) + \int \frac{\delta \mathcal{L}}{\delta q}(q)(\theta)(p_q - q)(\theta)\mathrm{d}\theta + \lambda \mathrm{KL}(p_q\|q) \\
&= \mathcal{L}(q) + \lambda \int \log \frac{q}{p_q}(\theta)(p_q - q)(\theta)\mathrm{d}\theta + \lambda \mathrm{KL}(p_q\|q) \\
&= \mathcal{L}(q) - \lambda \mathrm{KL}(q\|p_q).
\end{aligned}
$$

Moreover, by (16) with $q = q_*$ and the optimality condition $\frac{\delta \mathcal{L}}{\delta q}(q_*) = 0$, we get $\mathcal{L}(q) - \mathcal{L}(q_*) \ge \lambda \mathrm{KL}(q\|q_*)$. This finishes the proof. $\square$

We remark that the inequality (16) indicates that the functional $\mathcal{L}$ satisfies an analog of strong convexity with the proximal functional $\mathrm{KL}(q'\|q)$, and in the third

statement in Proposition 1, this convexity plays a similar role as in finite dimensional convex analysis. In particular, inequalities $\lambda\mathrm{KL}(q\|p_q) \geq \mathcal{L}(q) - \mathcal{L}(q_*)$ and $\mathcal{L}(q) - \mathcal{L}(q_*) \geq \lambda\mathrm{KL}(q\|q_*)$ can be recognized as the counterparts of the Polyak-Łojasiewicz and quadratic growth inequalities, respectively (for details of these conditions see Charles and Papailiopoulos (2018)). Following this analogy, $\mathrm{KL}(q\|p_q)$ and $\mathrm{KL}(q\|q_*)$ act as the squared norm of gradient at q and squared distance between q and q_*, respectively.

Finally, the third statement in Proposition 1 indicates that the divergence between q and p_q indeed measures the optimality gap, as expected from the optimality condition (15). Furthermore, we reveal that p_q can be interpreted as a *proximal point* which minimizes the sum of linearization of $\mathcal{L}$ and the KL-divergence around q, and that convergence of the optimization gap implies convergence to q_* in the sense of KL-divergence.

3.2 Convergence Rate in Continuous Time

We now introduce the convergence rate analysis by utilizing the aforementioned results. We first show that the mean field Langevin dynamics (8) converges linearly to the optimal solution of (4) in continuous time under the log-Sobolev inequality.

Theorem 1. *Let $\{q_t\}_{t\geq 0}$ be the evolution described by (8). Under Assumption 1, 2, we get for $t \geq 0$,*

$$\mathcal{L}(q_t) - \mathcal{L}(q_*) \leq \exp(-2\alpha\lambda t)(\mathcal{L}(q_0) - \mathcal{L}(q_*)).$$

Proof. From Proposition 1 and (9), we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{L}(q_t) - \mathcal{L}(q_*))$$
$$= \int \frac{\delta\mathcal{L}}{\delta q}(q_t)(\theta)\frac{\partial q_t}{\partial t}(\theta)\mathrm{d}\theta$$
$$= \lambda\int \frac{\delta\mathcal{L}}{\delta q}(q_t)(\theta)\nabla\cdot\left(q_t(\theta)\nabla\log\frac{q_t}{p_{q_t}}(\theta)\right)\mathrm{d}\theta$$
$$= -\lambda\int q_t(\theta)\nabla\frac{\delta\mathcal{L}}{\delta q}(q_t)(\theta)^\top\nabla\log\frac{q_t}{p_{q_t}}(\theta)\mathrm{d}\theta$$
$$= -\lambda^2\int q_t(\theta)\left\|\nabla\log\frac{q_t}{p_{q_t}}(\theta)\right\|_2^2\mathrm{d}\theta$$
$$\leq -2\alpha\lambda^2\mathrm{KL}(q_t\|p_{q_t})$$
$$\leq -2\alpha\lambda(\mathcal{L}(q_t) - \mathcal{L}(q_*)).$$

The statement then follows from a straightforward application of the Grönwall's inequality. $\square$

As a corollary, we can also show the convergence of $\mathrm{KL}(q_t\|p_{q_t})$ in the following sense.

Corollary 1. *Under the same setting as Theorem 1, we have for $t \geq 1$,*

$$\inf_{s\in[0,t]}\mathrm{KL}(q_s\|p_{q_s}) \leq \frac{\exp(-2\alpha\lambda(t-1))}{2\alpha\lambda^2}(\mathcal{L}(q_0) - \mathcal{L}(q_*)).$$

3.3 Convergence Rate in Discrete Time

For the standard Langevin dynamics (i.e., linear Fokker-Planck equation), exponential convergence in continuous time often implies the same convergence up to certain error (depending on the step size) in discrete time. In this section we show that the same property also holds for the mean field Langevin dynamics (7). We provide a convergence rate analysis of the discrete-time dynamics (10) by adapting a version of the "one-step interpolation" argument presented in Vempala and Wibisono (2019). Since analysis of one single step of the dynamics (10) is the key, we adopt the following notations for conciseness. Let $\theta^{(k)} \sim q^{(k)}(\theta)\mathrm{d}\theta$ be a random variable that represents the current iteration, and let $\theta_t^{(k+1)}$ be the next iterate of noisy gradient descent with step size $t > 0$:

$$\theta_t^{(k+1)} = \theta^{(k)} - t\nabla\frac{\delta F}{\delta q}(q^{(k)})(\theta^{(k)}) + \sqrt{2\lambda t}\xi^{(k)}, \quad (18)$$

where $\xi^{(k)} \sim \mathcal{N}(0, I_d)$. We denote a probability distribution of $\theta_t^{(k+1)}$ by $q_t^{(k+1)}(\theta)\mathrm{d}\theta$. Note that this step is equivalent to (10) when $t = \eta$, that is, $\theta_\eta^{(k+1)} = \theta^{(k+1)}$ and $q_\eta^{(k+1)} = q^{(k+1)}$. We define $\delta_{q^{(k)},t}$ as

$$\mathbb{E}_{(\theta^{(k)},\theta_t^{(k+1)})}\left\|\nabla\frac{\delta F}{\delta q}(q^{(k)})(\theta^{(k)}) - \nabla\frac{\delta F}{\delta q}(q_t^{(k+1)})(\theta_t^{(k+1)})\right\|_2^2,$$

where $\mathbb{E}_{(\theta^{(k)},\theta_t^{(k+1)})}$ is an expectation in the joint distribution of $\theta^{(k)}$ and $\theta_t^{(k+1)}$. Note that the term $\delta_{q^{(k)},t}$ can be recognized as a discretization error which comes from the positive step size $t > 0$, and this error usually decays to zero in common settings as $t \to 0$. Thus, by carefully incorporating this error into the proof of Theorem 1, we can show the convergence of the discrete-time (10) up to a certain error in proportion to a constant δ_η depending on η. The complete proof is deferred to Appendix A. 3.

Theorem 2. *Let $\{\theta^{(k)}\}_{k=0}^\infty$ be the iterations of random variables generated by the discrete-time dynamics (10) with the step size η and $\{q^{(k)}\}_{k=0}^\infty$ be the corresponding probability distributions. Suppose Assumption 1, 2 hold and there exists a constant δ_η such that $\delta_{q^{(k)},t} \leq \delta_\eta$ for any $0 < t \leq \eta$ and non-negative integer k. Then, it follows that*

$$\mathcal{L}(q^{(k)}) - \mathcal{L}(q_*) \leq$$
$$\frac{\delta_\eta}{2\alpha\lambda} + \exp(-\alpha\lambda\eta k)\left(\mathcal{L}(q^{(0)}) - \mathcal{L}(q_*)\right).$$

Proof sketch. We first present the one-step analysis for iteration (18) with step size η. Consider the stochastic differential equation:

$$\mathrm{d}\theta_t = -\nabla \frac{\delta F}{\delta q}(q^{(k)})(\theta_0)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}W_t, \qquad (19)$$

where $\theta_0 = \theta^{(k)}$ and W_t is the Brownian motion in $\mathbb{R}^d$ with $W_0 = 0$. Then, (18) is the solution of this equation at time t. We denote by $q_{0t}(\theta_0, \theta_t)$ the joint probability distribution of (θ_0, θ_t) for time t, and by $q_{t|0}$, $q_{0|t}$ and q_0, q_t conditional and marginal distributions. That is, $q_0 = q^{(k)}$, $q_t = q_t^{(k+1)}$ (i.e., $\theta_t \stackrel{\mathrm{d}}{=} \theta_t^{(k+1)}$), and

$$q_{0t}(\theta_0, \theta_t) = q_0(\theta_0)q_{t|0}(\theta_t|\theta_0) = q_t(\theta_t)q_{0|t}(\theta_0|\theta_t).$$

The continuity equation of $q_{t|0}$ conditioning on θ_0 can be described as follows (see Section 7 of Vempala and Wibisono (2019) for details):

$$\frac{\partial q_{t|0}(\theta_t|\theta_0)}{\partial t} = \nabla \cdot \left(q_{t|0}(\theta_t|\theta_0)\nabla \frac{\delta F}{\delta q}(q_0)(\theta_0) \right) + \lambda \Delta q_{t|0}(\theta_t|\theta_0).$$

Therefore, we obtain the following description of q_t:

$$\frac{\partial q_t(\theta_t)}{\partial t} = \lambda \nabla \cdot \left(q_t(\theta_t)\nabla \log \frac{q_t}{p_{q_t}}(\theta_t) \right)$$
$$+ \nabla \cdot \left\{ q_t(\theta_t)\left(\mathbb{E}_{\theta_0|\theta_t}\left[\nabla \frac{\delta F}{\delta q}(q_0)(\theta_0) \Big| \theta_t \right] \right.\right.$$
$$\left.\left. - \nabla \frac{\delta F}{\delta q}(q_t)(\theta_t) \right) \right\}, \qquad (20)$$

where $p_{q_t}(\cdot) \propto \exp\left(-\frac{1}{\lambda}\nabla \frac{\delta F}{\delta q}(q_t)(\cdot) \right)$. By Assumption 2 and (20), for $0 \le t \le \eta$, we have

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}t}(q_t) = \int \frac{\delta \mathcal{L}}{\delta q}(q_t)(\theta)\frac{\partial q_t}{\partial t}(\theta)\mathrm{d}\theta$$
$$\le -\frac{\lambda^2}{2}\int q_t(\theta)\left\| \nabla \log \frac{q_t}{p_{q_t}}(\theta) \right\|_2^2 \mathrm{d}\theta$$
$$+ \frac{1}{2}\mathbb{E}_{(\theta_0,\theta)\sim q_{0t}}\left\| \nabla \frac{\delta F}{\delta q}(q_0)(\theta_0) - \nabla \frac{\delta F}{\delta q}(q_t)(\theta) \right\|_2^2$$
$$\le -\alpha\lambda(\mathcal{L}(q_t) - \mathcal{L}(q_*)) + \frac{1}{2}\delta_\eta,$$

where we used $(\theta_0, \theta_t) \stackrel{\mathrm{d}}{=} (\theta^{(k)}, \theta_t^{(k+1)})$ to bound the last expectation by $\delta_{q^{(k)},t} \le \delta_\eta$. Noting $q_\eta = q^{(k+1)}$ and $q_0 = q^{(k)}$, Grönwall's inequality yields

$$\mathcal{L}(q^{(k+1)}) - \mathcal{L}(q_*) - \frac{\delta_\eta}{2\alpha\lambda}$$
$$\le \exp(-\alpha\lambda\eta)\left(\mathcal{L}(q^{(k)}) - \mathcal{L}(q_*) - \frac{\delta_\eta}{2\alpha\lambda} \right).$$

This reduction holds at every iteration of (19), which concludes the proof. $\qquad \square$

The proof indicates that the continuous equation (20) of the dynamics (19) associated with noisy gradient descent contains a discretization error term compared to the continuous-time counterpart (7). Usually, this error depends on the step size η, hence η should be chosen sufficiently small to achieve a required optimization accuracy. Moreover, to derive a convergence rate, the relationship between η and the discretization error should be explicitly characterized. The following lemma outlines this dependency in the case of mean field neural networks (13).

Lemma 1. *Consider the loss minimization setting using mean field neural networks in Section 2.3 with $r(\theta) = \|\theta\|_2^2$. Suppose that $\ell(\cdot, y)$ and $h(\cdot, x)$ are differentiable and sufficiently smooth, that is, there exist positive constants $C_1, \ldots, C_4$ such that $|\partial_z \ell(z, y)| \le C_1$, $|\partial_z \ell(z, y) - \partial_z \ell(z', y)| \le C_2|z - z'|$, $\|\partial_\theta h_\theta(x)\|_2 \le C_3$, and $\|\partial_\theta h_\theta(x) - \partial_\theta h_{\theta'}(x)\|_2 \le C_4\|\theta - \theta'\|_2$. Also suppose $2\lambda'\eta < 1$ and the following condition holds for the iterate $\theta^{(k)} \sim q^{(k)}(\theta)\mathrm{d}\theta$.*

$$\mathbb{E}_{\theta^{(k)}}[\|\theta^{(k)}\|_2^2] \le \frac{\eta C_1^2 C_3^2 + 2\lambda d}{2\eta\lambda'^2}. \qquad (21)$$

Then, we get for any $0 < t \le \eta$,

$$\delta_{q^{(k)},t} \le 40\eta(C_2^2 C_3^4 + (C_1 C_4 + 2\lambda')^2)(\eta C_1^2 C_3^2 + \lambda d). \qquad (22)$$

In addition, the same bound as (21) also holds for the second moment of the next iterate $\|\theta^{(k+1)}\|_2$.

This is to say, $\delta_{q^{(k)},t} = O(\eta)$ for all k as long as (21) holds for $k = 0$. Therefore, in combination with Theorem 2, we arrive at a convergence rate guarantee for the discrete-time dynamics in optimizing mean field neural networks (up to certain error). Specifically, the following Corollary implies an iteration complexity of $O\left(\frac{1}{\epsilon\alpha^2\lambda^2}\log\frac{1}{\epsilon} \right)$ to achieve an ϵ-accurate solution.

Corollary 2. *Consider the same setting as Lemma 1 and suppose Assumption 2 holds. Let $\{\theta^{(k)}\}_{k=0}^\infty$ be the iterations of random variables generated by the discrete-time dynamics (10) with the step size $\eta = O(\epsilon\alpha\lambda)$ and $\{q^{(k)}\}_{k=0}^\infty$ be the corresponding probability distributions. Then if the condition (21) is satisfied for the initial iterate $\theta^{(0)} \sim q^{(0)}(\theta)\mathrm{d}\theta$, we know that for any step size $2\lambda'\eta < 1$, the following statement holds true for $k = 0, 1, 2, 3...$,*

$$\mathcal{L}(q^{(k)}) - \mathcal{L}(q_*) =$$
$$O(\epsilon) + \exp\left(-O(\epsilon\alpha^2\lambda^2 k) \right)\left(\mathcal{L}(q^{(0)}) - \mathcal{L}(q_*) \right). \quad (23)$$

Finally, we remark that discretization error induced by finite-particle approximation can also be controlled via a direct application of Theorem 3 of Mei et al. (2018). However, such finite-particle error grows exponentially

with the time horizon, and thus is not negligible unless the exponent in the linear convergence (23) is sufficiently large. In future work, we intend to investigate conditions under which such exponential dependence can be avoided (e.g., as in Chen et al. (2020b)).

4 PRIMAL-DUAL VIEWPOINT

As seen in Section 3, the proximal Gibbs distribution p_q plays an important role in our convergence rate analysis. In this section, we complement the previous results by presenting a primal and dual perspective of this proximal distribution in the (regularized) empirical risk minimization setting. Based on this connection, we show that the duality gap can be minimized by the mean field Langevin dynamics (7).

4.1 Primal-dual Problem

We first introduce a primal-dual formulation of the empirical risk minimization problem. For a train dataset $\{(x_i, y_i)\}_{i=1}^n$ and differentiable convex loss function: $\ell(z, y)$ in z, we consider the minimization of the following regularized empirical risk:

$$\mathcal{L}(q) = \frac{1}{n}\sum_{i=1}^n \ell(h_q(x_i), y_i) + \lambda' \mathbb{E}_{\theta \sim q}[\|\theta\|_2^2] + \lambda \mathbb{E}_q[\log q].$$
(24)

Note that this problem is a special case of (4) by setting $F(q) = \frac{1}{n}\sum_{i=1}^n \ell(h_q(x_i), y_i) + \lambda' \mathbb{E}_{\theta \sim q}[\|\theta\|_2^2]$. Write $\ell_i(z) = \ell(z, y_i)$ and $\ell_i^*(\cdot)$ as its Fenchel conjugate, i.e.,

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

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

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

Then, the dual problem of (24) is defined as

$$\max_{g \in \mathbb{R}^n}\left\{\mathcal{D}(g) = -\frac{1}{n}\sum_{i=1}^n \ell_i^*(g_i) - \lambda \log \int q_g(\theta)\mathrm{d}\theta\right\}.$$
(25)

The duality theorem (Rockafellar, 1970; Bauschke et al., 2011; Oko et al., 2022) guarantees the relationship $\mathcal{D}(g) \leq \mathcal{L}(q)$ for any $g \in \mathbb{R}^n$ and $q \in \mathcal{P}$, and it is known that the duality gap $\mathcal{L}(q) - \mathcal{D}(g)$ vanishes at the solutions of (24) and (25) when they exist. In our problem setting, it is possible to establish a stronger and more precise result.

We denote $g_q = \{\partial_z \ell(z, y_i)|_{z=h_q(x_i)}\}_{i=1}^n \in \mathbb{R}^n$ ($q \in \mathcal{P}$). The following theorem exactly characterizes the duality gap $\mathcal{L}(q) - \mathcal{D}(g_q)$ between $q \in \mathcal{P}$ and $g_q \in \mathbb{R}^n$ via the proximal Gibbs distribution p_q.

Theorem 3 (Duality Theorem). *Suppose $\ell(\cdot, y)$ is convex and differentiable. For any $q \in \mathcal{P}$ the duality gap between $q \in \mathcal{P}$ and g_q of the problems (24) and (25) is*

$$0 \leq \mathcal{L}(q) - \mathcal{D}(g_q) = \lambda \mathrm{KL}(q\|p_q).$$

We make the following observations. First, this theorem can be seen as a refinement of Proposition 1, in which $\lambda \mathrm{KL}(q\|p_q)$ upper bounds the optimality gap $\mathcal{L}(q) - \mathcal{L}(q_*)$ in the general setting. Theorem 3 reveals that $\lambda \mathrm{KL}(q\|p_q)$ is *exactly* the duality gap $\mathcal{L}(q) - \mathcal{D}(g_q)$ for the empirical risk minimization problems. Second, because of the relationship $p_q(\theta) = \frac{q_{g_q}(\theta)}{\int q_{g_q}(\theta)\mathrm{d}\theta}$, we notice that the proximal Gibbs distribution p_q of q can be seen as a "round trip" between the primal and dual spaces: $q \to g_q \to q_{g_q} \propto p_q$; the equation $\mathcal{L}(q) - \mathcal{D}(g_q) = \lambda \mathrm{KL}(q\|p_q)$ gives an interesting relationship among these variables. Third, Combining Theorem 3 and the duality theorem (e.g., see Proposition 1 of Oko et al. (2022)): $\mathcal{D}(g) \leq \mathcal{L}(q)$ for any $q \in \mathcal{P}$ and $g \in \mathbb{R}^n$, we see that g_{q_*} is the solution of the dual problem (25).

Another interesting quantity to investigate is the primal objective $\mathcal{L}(p_q)$ of the proximal Gibbs distribution p_q. The next theorem gives an upper bound on a duality gap between p_q and a dual variable g_q.

Theorem 4 (Second Duality Theorem). *Suppose $\ell(\cdot, y)$ is convex and C_2-smooth, that is, $|\partial_z \ell(z, y) - \partial_z \ell(z', y)| \leq C_2|z - z'|$, and $|h_\theta(x)| \leq B$. Then for any $q \in \mathcal{P}$, the duality gap between $p_q \in \mathcal{P}$ and g_q of the problems (24) and (25) satisfies*

$$0 \leq \mathcal{L}(p_q) - \mathcal{D}(g_q) \leq (\lambda + 2B^2 C_2)\mathrm{KL}(q\|p_q).$$

This theorem provides a new choice for solving the problem (24). That is, after obtaining q by the optimization, we can alternatively utilize p_q as an alternative solution, which may be efficiently approximated by sampling methods for Gibbs distributions.

4.2 Convergence of Duality Gap

Combining Theorem 3 with Corollary 1, we see that mean field Langevin dynamics solves both the primal and dual problems in the following sense.

Corollary 3. *Consider the evolution $\{q_t\}_{t \geq 0}$, which satisfies (8) for the problem (24) under the same settings as in Corollary 1 and Theorem 3. Let $\{g_{q_t}\}_{t \geq 0}$ be the associated dynamics in $\mathbb{R}^n$. Then for $t \geq 1$,*

$$\inf_{s \in [0,t]}\{\mathcal{L}(q_s) - \mathcal{D}(g_{g_s})\}$$

$$\leq \frac{\exp(-2\alpha\lambda(t-1))}{2\alpha\lambda}(\mathcal{L}(q_0) - \mathcal{L}(q_*)).$$

Atsushi Nitanda, Denny Wu, and Taiji Suzuki

Thus, we can conclude that mean field Langevin dynamics also solves the dual problem, that is, $\mathcal{D}(g_{q_t})$ also converges to $\mathcal{D}(g_{q_*})$. Following the same reasoning, we can derive a corollary of Theorem 4 demonstrating the convergence of $\mathcal{L}(p_{q_t}) - \mathcal{D}(g_{g_t})$.

Evaluation of duality gap. One benefit of this primal-dual formulation is that we can observe the optimization gap by computing the duality gap $\mathcal{L}(q) - \mathcal{D}(g)$ or $\mathcal{L}(p_q) - \mathcal{D}(g)$, without the knowledge of the optimal value $\mathcal{L}(q_*)$. In Figure 2 we empirically demonstrate the duality gap on a regression problem with the squared loss. We set $n = 1000, d = 5$, and consider a simple student-teacher setting, where the teacher model is a two-layer sigmoid network with orthogonal neurons, and the student model h_Θ is a two-layer mean field neural network of width $M = 1000$ with tanh activation. The student model is optimized by the noisy gradient descent with $\eta = 0.01$, and we use the Langevin algorithm to obtain approximate samples from the proximal Gibbs distribution p_q. For the primal objective $\mathcal{L}$, we adopt the k-nearest neighbors estimator (Kozachenko and Leonenko, 1987) with $k = 10$ to estimate the entropy; for the dual objective $\mathcal{D}$, the approximation of the log integral term is described in Appendix B. 1.
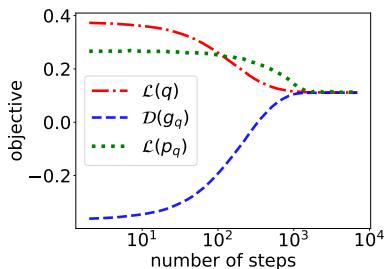

Figure 2: Illustration of duality gap: two-layer tanh network optimizing the empirical squared error. We set $\lambda = \lambda' = 10^{-2}$.

Observe that towards the end of training, the primal ($\mathcal{L}(q)$ and $\mathcal{L}(p_q)$) and dual ($\mathcal{D}(g_q)$) objectives become close, which is consistent with Theorem 3 and 4.

Conclusion

We established quantitative global convergence guarantee for the mean field Langevin dynamics in both continuous- and discrete-time, by adapting convex optimization techniques into the space of measures, in combination with standard analysis of the Langevin dynamics. Looking forward, an interesting future direction is to conduct the analysis under weaker isoperimetry conditions such as the Poincaré inequality, which covers more general objectives. It is also important to refine our convergence result (e.g., exponential dependence on $1/\lambda$ in the LSI constant) under additional structure of the learning problem. Another interesting direction is to explore applications of the mean field dynamics beyond the optimization of neural networks.

References

Akiyama, S. and Suzuki, T. (2021). On learnability via gradient method for two-layer relu neural networks in teacher-student setting. In *Proceedings of International Conference on Machine Learning 38*, pages 152–162.

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

Bakry, D., Gentil, I., and Ledoux, M. (2013). *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media.

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

Bou-Rabee, N. and Eberle, A. (2021). Mixing time guarantees for unadjusted hamiltonian monte carlo. *arXiv e-prints*, pages arXiv–2105.

Bou-Rabee, N. and Schuh, K. (2020). Convergence of unadjusted hamiltonian monte carlo for mean-field models. *arXiv preprint arXiv:2009.08735*.

Charles, Z. and Papailiopoulos, D. (2018). Stability and generalization of learning algorithms that converge to global optima. In *Proceedings of International Conference on Machine Learning 35*, pages 745–754.

Chen, Z., Cao, Y., Gu, Q., and Zhang, T. (2020a). A generalized neural tangent kernel analysis for two-layer neural networks. *arXiv preprint arXiv:2002.04026*.

Chen, Z., Rotskoff, G. M., Bruna, J., and Vanden-Eijnden, E. (2020b). A dynamical central limit theorem for shallow neural networks. *arXiv preprint arXiv:2008.09623*.

Chizat, L. (2021a). Convergence rates of gradient methods for convex optimization in the space of measures. *arXiv preprint arXiv:2105.08368*.

Chizat, L. (2021b). Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, pages 1–46.

Chizat, L. (2022). Mean-field langevin dynamics: Exponential convergence and annealing. *arXiv preprint arXiv:2202.01009*.

Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized

models using optimal transport. In *Advances in Neural Information Processing Systems 31*, pages 3040–3050.

Dalalyan, A. S. (2014). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *arXiv preprint arXiv:1412.7392*.

Dalalyan, A. S. (2017). Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. *arXiv preprint arXiv:1704.04752*.

Durmus, A. and Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587.

Eberle, A., Guillin, A., and Zimmer, R. (2019). Quantitative harris-type theorems for diffusions and mckean–vlasov processes. *Transactions of the American Mathematical Society*, 371(10):7135–7173.

Erdogdu, M. A. and Hosseinzadeh, R. (2020). On the convergence of langevin monte carlo: The interplay between tail growth and smoothness. *arXiv preprint arXiv:2005.13097*.

Erdogdu, M. A., Hosseinzadeh, R., and Zhang, M. S. (2021). Convergence of langevin monte carlo in chi-squared and rényi divergence.

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

Guillin, A., Liu, W., Wu, L., and Zhang, C. (2019). Uniform poincar {\'e} and logarithmic sobolev inequalities for mean field particles systems. *arXiv preprint arXiv:1909.07051*.

Guillin, A., Liu, W., Wu, L., and Zhang, C. (2021). The kinetic fokker-planck equation with mean field interaction. *Journal de Mathématiques Pures et Appliquées*, 150:1–23.

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

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

Jabir, J.-F., Šiška, D., and Szpruch, Ł. (2019). Mean-field neural odes via relaxed optimal control. *arXiv preprint arXiv:1912.05475*.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, pages 8580–8589.

Javanmard, A., Mondelli, M., and Montanari, A. (2019). Analysis of a two-layer neural network via displacement convexity. *arXiv preprint arXiv:1901.01375*.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17.

Kazeykina, A., Ren, Z., Tan, X., and Yang, J. (2020). Ergodicity of the underdamped mean-field langevin dynamics. *arXiv preprint arXiv:2007.14660*.

Kent, C., Blanchet, J., and Glynn, P. (2021). Frank-wolfe methods in probability space. *arXiv preprint arXiv:2105.05352*.

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

Li, X., Wu, Y., Mackey, L., and Erdogdu, M. A. (2019). Stochastic runge-kutta accelerates langevin monte carlo and beyond. In *Advances in Neural Information Processing Systems 32*, pages 7748–7760.

Mei, S., Misiakiewicz, T., and Montanari, A. (2019). Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *arXiv preprint arXiv:1902.06015*.

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

Menz, G. and Schlichting, A. (2014). Poincaré and logarithmic sobolev inequalities by decomposition of the energy landscape. *The Annals of Probability*, 42(5):1809–1884.

Milstein, G. N. and Tretyakov, M. V. (2013). *Stochastic numerics for mathematical physics*. Springer Science & Business Media.

Monmarché, P. (2017). Long-time behaviour and propagation of chaos for mean field kinetic particles. *Stochastic Processes and their Applications*, 127(6):1721–1737.

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

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

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

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

Rotskoff, G. M., Jelassi, S., Bruna, J., and Vanden-Eijnden, E. (2019). Global convergence of neuron birth-death dynamics. In *Proceedings of International Conference on Machine Learning 36*, pages 9689–9698.

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

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

Suzuki, T. (2019). Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *Proceedings of the 7th International Conference on Learning Representations*.

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

Villani, C. (2009). Hypocoercivity.

Wei, C., Lee, J. D., Liu, Q., and Ma, T. (2019). Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems 32*, pages 9712–9724.

Wibisono, A. (2018). Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Proceedings of Conference on Learning Theory 31*, pages 2093–3027.

Ying, L. (2020). Mirror descent algorithms for minimizing interacting free energy. *Journal of Scientific Computing*, 84(3):1–14.

Appendix: Convex Analysis of Mean Field Langevin Dynamics

A   OMITTED PROOFS

A. 1   Preliminaries: Log-Sobolev Inequality

For the proximal Gibbs distribution (14) of the mean field neural network, we can verify that the log-Sobolev inequality (Assumption 2) holds *uniformly* over q under the boundedness assumptions on $\partial_z \ell(z, y)$ and $h_\theta(x)$ and ℓ_2-regularization $r(\theta) = \|\theta\|_2^2$ by utilizing the following two facts. Specifically, these two facts indicate that the Gibbs distribution p_q whose potential is the sum of strongly concave function and bounded perturbation satisfies the log-Sobolev inequality.

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

Example A (Bakry and 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 $\nabla^2 f \succeq c I_p$, then $q(\theta)\mathrm{d}\theta$ satisfies Log-Sobolev inequality with constant c.*

In addition, the log-Sobolev inequality is preserved under bounded perturbation, as originally shown in Holley and Stroock (1987).

Lemma A (Holley and 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 $\alpha/\exp(4\|B\|_\infty)$.

In our case, if $|\partial_z \ell(z, y)| \le C_1$, $|h_\theta(x)| \le C_5$, and $r(\theta) = \|\theta\|_2^2$, then the proximal Gibbs distribution (14) satisfies the log-Sobolev inequality with a constant $\alpha = \frac{2\lambda'}{\lambda \exp(4C_1 C_5 \lambda^{-1})}$. We remark that the exponential dependence in the LSI constant may be unavoidable in the most general setting (Menz and Schlichting, 2014).

A. 2   Continuous Time Analysis

Proof of Corollary 1. In the same way as the proof of Theorem 1,

$$2\alpha\lambda^2 \mathrm{KL}(q_s\|p_{q_s}) \le -\frac{\mathrm{d}}{\mathrm{d}s}(\mathcal{L}(q_s) - \mathcal{L}(q_*)).$$

By taking integral of this inequality on the interval $[t-1, t]$ ($t \in \mathbb{N}$), we get

$$2\alpha\lambda^2 \int_{t-1}^t \mathrm{KL}(q_s\|p_{q_s})\mathrm{d}s \le \mathcal{L}(q_{t-1}) - \mathcal{L}(q_*) - (\mathcal{L}(q_t) - \mathcal{L}(q_*))$$

$$\le \exp(-2\alpha\lambda(t-1))(\mathcal{L}(q_0) - \mathcal{L}(q_*)).$$

Therefore, there exists $s_t \in [t-1, t]$ ($t \in \mathbb{N}$) such that

$$\mathrm{KL}(q_{s_t}\|p_{q_{s_t}}) \le \frac{\exp(-2\alpha\lambda(t-1))}{2\alpha\lambda^2}(\mathcal{L}(q_0) - \mathcal{L}(q_*)).$$

By taking the infimum of $\mathrm{KL}(q_s\|p_{q_s})$ over $[0, t]$ ($t \in \mathbb{N}$), we have

$$\inf_{s \in [0,t]} \mathrm{KL}(q_s\|p_{q_s}) \le \frac{\exp(-2\alpha\lambda(t-1))}{2\alpha\lambda^2}(\mathcal{L}(q_0) - \mathcal{L}(q_*)).$$

$\square$

A. 3 Discrete Time Analysis

Proof of Theorem 2. We first present the one step analysis for the iteration (18) with the step size η. Let us consider the stochastic differential equation:

$$\mathrm{d}\theta_t = -\nabla \frac{\delta F}{\delta q}(q^{(k)})(\theta_0)\mathrm{d}t + \sqrt{2\lambda}\mathrm{d}W_t, \tag{26}$$

where $\theta_0 = \theta^{(k)}$ and W_t is the Brownian motion in $\mathbb{R}^d$ with $W_0 = 0$. Then, the step (18) is the solution of this equation at time t. We denote by $q_{0t}(\theta_0, \theta_t)$ the joint probability distribution of (θ_0, θ_t) for time t, and by $q_{t|0}$, $q_{0|t}$ and q_0, q_t conditional and marginal distributions. That is, it holds that $q_0 = q^{(k)}$, $q_t = q_t^{(k+1)}$ (i.e., $\theta_t \overset{d}{=} \theta_t^{(k+1)}$), and

$$q_{0t}(\theta_0, \theta_t) = q_0(\theta_0)q_{t|0}(\theta_t|\theta_0) = q_t(\theta_t)q_{0|t}(\theta_0|\theta_t).$$

The continuity equation of $q_{t|0}$ conditioned on θ_0 is given as (see Section 7 of Vempala and Wibisono (2019) for details):

$$\frac{\partial q_{t|0}(\theta_t|\theta_0)}{\partial t} = \nabla \cdot \left(q_{t|0}(\theta_t|\theta_0)\nabla \frac{\delta F}{\delta q}(q_0)(\theta_0) \right) + \lambda \Delta q_{t|0}(\theta_t|\theta_0).$$

Therefore, we obtain the continuity equation of q_t:

$$\begin{aligned}
\frac{\partial q_t(\theta_t)}{\partial t} &= \int \frac{\partial q_{t|0}(\theta_t|\theta_0)}{\partial t} q_0(\theta_0)\mathrm{d}\theta_0 \\
&= \int \left(\nabla \cdot \left(q_{0t}(\theta_0, \theta_t)\nabla \frac{\delta F}{\delta q}(q_0)(\theta_0) \right) + \lambda \Delta q_{0t}(\theta_0, \theta_t) \right) \mathrm{d}\theta_0 \\
&= \nabla \cdot \left(q_t(\theta_t)\int q_{0|t}(\theta_0|\theta_t)\nabla \frac{\delta F}{\delta q}(q_0)(\theta_0)\mathrm{d}\theta_0 \right) + \lambda \Delta q_t(\theta_t) \\
&= \nabla \cdot \left(q_t(\theta_t)\left(\mathbb{E}_{\theta_0|\theta_t}\left[\nabla \frac{\delta F}{\delta q}(q_0)(\theta_0)\Big|\theta_t \right] + \lambda \nabla \log q_t(\theta_t) \right) \right) \\
&= \lambda \nabla \cdot \left(q_t(\theta_t)\nabla \log \frac{q_t}{p_{q_t}}(\theta_t) \right) \\
&\quad + \nabla \cdot \left(q_t(\theta_t)\left(\mathbb{E}_{\theta_0|\theta_t}\left[\nabla \frac{\delta F}{\delta q}(q_0)(\theta_0)\Big|\theta_t \right] - \nabla \frac{\delta F}{\delta q}(q_t)(\theta_t) \right) \right),
\end{aligned} \tag{27}$$

where $p_{q_t}(\cdot) \propto \exp\left(-\frac{1}{\lambda}\nabla \frac{\delta F}{\delta q}(q_t)(\cdot) \right)$. For simplicity, we write $\delta_t(\cdot) = \mathbb{E}_{\theta_0 \sim q_{0|t}}\left[\nabla \frac{\delta F}{\delta q}(q_0)(\theta_0)\Big|\theta_t = \cdot \right] - \nabla \frac{\delta F}{\delta q}(q_t)(\cdot)$. By Assumption 2 and (27), for $0 \le t \le \eta$, we have

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}t}(q_t) &= \int \frac{\delta \mathcal{L}}{\delta q}(q_t)(\theta)\frac{\partial q_t}{\partial t}(\theta)\mathrm{d}\theta \\
&= \lambda \int \frac{\delta \mathcal{L}}{\delta q}(q_t)(\theta)\nabla \cdot \left(q_t(\theta)\nabla \log \frac{q_t}{p_{q_t}}(\theta) \right) \mathrm{d}\theta \\
&\quad + \int \frac{\delta \mathcal{L}}{\delta q}(q_t)(\theta)\nabla \cdot (q_t(\theta)\delta_t(\theta)) \mathrm{d}\theta \\
&= -\lambda \int q_t(\theta)\nabla \frac{\delta \mathcal{L}}{\delta q}(q_t)(\theta)^\top \nabla \log \frac{q_t}{p_{q_t}}(\theta)\mathrm{d}\theta \\
&\quad - \int q_t(\theta)\nabla \frac{\delta \mathcal{L}}{\delta q}(q_t)(\theta)^\top \delta_t(\theta)\mathrm{d}\theta \\
&= -\lambda^2 \int q_t(\theta)\left\| \nabla \log \frac{q_t}{p_{q_t}}(\theta) \right\|_2^2 \mathrm{d}\theta \\
&\quad - \int q_{0t}(\theta_0, \theta)\lambda \nabla \log \frac{q_t}{p_{q_t}}(\theta)^\top \left(\nabla \frac{\delta F}{\delta q}(q_0)(\theta_0) - \nabla \frac{\delta F}{\delta q}(q_t)(\theta) \right) \mathrm{d}\theta_0\mathrm{d}\theta \\
&\le -\lambda^2 \int q_t(\theta)\left\| \nabla \log \frac{q_t}{p_{q_t}}(\theta) \right\|_2^2 \mathrm{d}\theta
\end{aligned}$$

$$+ \frac{1}{2} \int q_{0t}(\theta_0, \theta) \left(\lambda^2 \left\| \nabla \log \frac{q_t}{p_{q_t}}(\theta) \right\|_2^2 + \left\| \nabla \frac{\delta F}{\delta q}(q_0)(\theta_0) - \nabla \frac{\delta F}{\delta q}(q_t)(\theta) \right\|_2^2 \right) \mathrm{d}\theta_0 \mathrm{d}\theta$$

$$\leq -\frac{\lambda^2}{2} \int q_t(\theta) \left\| \nabla \log \frac{q_t}{p_{q_t}}(\theta) \right\|_2^2 \mathrm{d}\theta$$

$$+ \frac{1}{2} \mathbb{E}_{(\theta_0, \theta) \sim q_{0t}} \left[\left\| \nabla \frac{\delta F}{\delta q}(q_0)(\theta_0) - \nabla \frac{\delta F}{\delta q}(q_t)(\theta) \right\|_2^2 \right]$$

$$\leq -\alpha \lambda^2 \mathrm{KL}(q_t \| p_{q_t}) + \frac{1}{2} \delta_{q_0, t}$$

$$\leq -\alpha \lambda (\mathcal{L}(q_t) - \mathcal{L}(q_*)) + \frac{1}{2} \delta_\eta,$$

where we used $(\theta_0, \theta_t) \stackrel{\mathrm{d}}{=} (\theta^{(k)}, \theta_t^{(k+1)})$ to bound the last expectation by $\delta_{q^{(k)}, t} \leq \delta_\eta$. Thus, for $0 \leq t \leq \eta$, we get

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\mathcal{L}(q_t) - \mathcal{L}(q_*) - \frac{\delta_\eta}{2\alpha\lambda} \right) \leq -\alpha\lambda \left(\mathcal{L}(q_t) - \mathcal{L}(q_*) - \frac{\delta_\eta}{2\alpha\lambda} \right).$$

Noting $q_\eta = q^{(k+1)}$ and $q_0 = q^{(k)}$, the Gronwall's inequality leads to

$$\mathcal{L}(q^{(k+1)}) - \mathcal{L}(q_*) - \frac{\delta_\eta}{2\alpha\lambda} \leq \exp(-\alpha\lambda\eta) \left(\mathcal{L}(q^{(k)}) - \mathcal{L}(q_*) - \frac{\delta_\eta}{2\alpha\lambda} \right).$$

This reduction holds at every iteration of (26). Hence, we arrive at the desired result,

$$\mathcal{L}(q^{(k)}) - \mathcal{L}(q_*) \leq \frac{\delta_\eta}{2\alpha\lambda} + \exp(-\alpha\lambda\eta k) \left(\mathcal{L}(q^{(0)}) - \mathcal{L}(q_*) - \frac{\delta_\eta}{2\alpha\lambda} \right).$$

□

Proof of Lemma 1. For notational simplicity, we use θ, θ', q, q' to represent $\theta^{(k)}, \theta_t^{(k+1)}, q^{(k)}, q_t^{(k+1)}$ appearing in $\delta_{q^{(k)}, t}$. Recall the definition of F in Section 2.3, we have

$$\left\| \nabla \frac{\delta F}{\delta q}(q)(\theta) - \nabla \frac{\delta F}{\delta q}(q')(\theta') \right\|_2 \leq \mathbb{E}_{(X,Y)} \left[\| \partial_z \ell(h_q(X), Y) \partial_\theta h_\theta(X) - \partial_z \ell(h_{q'}(X), Y) \partial_\theta h_{\theta'}(X) \|_2 \right]$$

$$+ 2\lambda' \|\theta - \theta'\|_2$$

$$\leq \mathbb{E}_{(X,Y)} \left[\|(\partial_z \ell(h_q(X), Y) - \partial_z \ell(h_{q'}(X), Y)) \partial_\theta h_\theta(X) \|_2 \right]$$

$$+ \mathbb{E}_{(X,Y)} \left[\| \partial_z \ell(h_{q'}(X), Y)(\partial_\theta h_\theta(X) - \partial_\theta h_{\theta'}(X)) \|_2 \right]$$

$$+ 2\lambda' \|\theta - \theta'\|_2$$

$$\leq C_2 C_3 \mathbb{E}_X [|h_q(X) - h_{q'}(X)|] + (C_1 C_4 + 2\lambda') \|\theta - \theta'\|_2. \tag{28}$$

The expectation of $\|\theta - \theta'\|_2^2$ can be bounded as follows:

$$\mathbb{E}_{(\theta, \theta')} \left[\|\theta - \theta'\|_2^2 \right] = \mathbb{E}_{(\theta, \xi)} \left[\left\| t \nabla \frac{\delta F}{\delta q}(q)(\theta) - \sqrt{2\lambda t} \xi \right\|_2^2 \right]$$

$$\leq 2t^2 \mathbb{E}_\theta \left[\left\| \nabla \frac{\delta F}{\delta q}(q)(\theta) \right\|_2^2 \right] + 4\lambda t \mathbb{E}_\xi \left[\|\xi\|_2^2 \right]$$

$$\leq 4t^2 \mathbb{E}_\theta \left[\| \mathbb{E}_{(X,Y)} [\partial_z \ell(h_q(X), Y) \partial_\theta h_\theta(X)] \|_2^2 + 4\lambda'^2 \|\theta\|_2^2 \right] + 4\lambda t d$$

$$\leq 4t^2 C_1^2 C_3^2 + 16t^2 \lambda'^2 \mathbb{E}_\theta \left[\|\theta\|_2^2 \right] + 4\lambda t d$$

$$\leq 20\eta (\eta C_1^2 C_3^2 + \lambda d). \tag{29}$$

Moreover, $|h_q(x) - h_{q'}(x)|$ can be bounded as follows:

$$
\begin{aligned}
|h_q(x) - h_{q'}(x)| &= \left| \mathbb{E}_{(\theta,\theta')}[h_\theta(x) - h_{\theta'}(x)] \right| \\
&\leq C_3 \mathbb{E}_{(\theta,\theta')}[\|\theta - \theta'\|_2] \\
&\leq C_3 \sqrt{\mathbb{E}_{(\theta,\theta')}[\|\theta - \theta'\|_2^2]}.
\end{aligned}
\tag{30}
$$

Therefore, we get by (28), (29), and (30),

$$
\begin{aligned}
\mathbb{E}_{(\theta,\theta')}\left[\left\| \nabla \frac{\delta F}{\delta q}(q)(\theta) - \nabla \frac{\delta F}{\delta q}(q')(\theta') \right\|_2^2 \right] &\leq 2C_2^2 C_3^2 (\mathbb{E}_X[|h_q(X) - h_{q'}(X)|])^2 \\
&\quad + 2(C_1 C_4 + 2\lambda')^2 \mathbb{E}_{(\theta,\theta')}\left[\|\theta - \theta'\|_2^2 \right] \\
&\leq 2(C_2^2 C_3^4 + (C_1 C_4 + 2\lambda')^2) \mathbb{E}_{(\theta,\theta')}\left[\|\theta - \theta'\|_2^2 \right] \\
&\leq 40\eta(C_2^2 C_3^4 + (C_1 C_4 + 2\lambda')^2)(\eta C_1^2 C_3^2 + \lambda d).
\end{aligned}
$$

Finally, we show the same bound on second moment of $\|\theta^{(k+1)}\|_2$ as on $\|\theta^{(k)}\|_2$. Using the inequality $(a+b)^2 \leq (1+\gamma)a^2 + \left(1 + \frac{1}{\gamma}\right)b^2$ with $\gamma = \frac{2\eta\lambda'}{1-2\eta\lambda'}$, we get

$$
\begin{aligned}
\mathbb{E}_{\theta^{(k+1)}}\left[\|\theta^{(k+1)}\|_2^2\right] &= \mathbb{E}_{(\theta,\xi)}\left[\left\| (1 - 2\lambda'\eta)\theta - \eta \mathbb{E}_{(X,Y)}[\partial_z \ell(h_q(X), Y)\partial_\theta h_\theta(X)] + \sqrt{2\lambda\eta}\xi \right\|_2^2 \right] \\
&\leq \mathbb{E}_{(\theta,\xi)}\left[\left((1 - 2\lambda'\eta)\|\theta\|_2 + \eta C_1 C_3 + \sqrt{2\lambda\eta}\|\xi\|_2 \right)^2 \right] \\
&= \mathbb{E}_{(\theta,\xi)}\left[(1+\gamma)(1 - 2\lambda'\eta)^2 \|\theta\|_2^2 + \left(1 + \frac{1}{\gamma}\right)\left(\eta C_1 C_3 + \sqrt{2\lambda\eta}\|\xi\|_2\right)^2 \right] \\
&\leq (1 - 2\lambda'\eta)\mathbb{E}_\theta\left[\|\theta\|_2^2\right] + \frac{1}{\lambda'}\left(\eta C_1^2 C_3^2 + 2\lambda\mathbb{E}_\xi\left[\|\xi\|_2^2\right] \right) \\
&\leq (1 - 2\lambda'\eta)\frac{\eta C_1^2 C_3^2 + 2\lambda d}{2\eta\lambda'^2} + \frac{1}{\lambda'}\left(\eta C_1^2 C_3^2 + 2\lambda d \right) \\
&= \frac{\eta C_1^2 C_3^2 + 2\lambda d}{2\eta\lambda'^2}.
\end{aligned}
$$

$\square$

A. 4 Duality Theorems

Proof of Theorem 3. It is well known that $\nabla \ell_i$ and $\nabla \ell_i^*$ give inverse mapping to each other. Hence, we know that $\nabla \ell_i^*(g_{q,i}) = \nabla \ell_i^*(\nabla \ell_i(h_q(x_i))) = h_q(x_i)$ and

$$
\ell_i^*(g_{q,i}) = \sup_{z \in \mathbb{R}}\{z g_{q,i} - \ell_i(z)\} = h_q(x_i)g_{q,i} - \ell_i(h_q(x_i)).
\tag{31}
$$

Recall the definitions of p_q, g_q, and q_g, we see that $q_{g_q} \propto p_q$, and hence

$$
\mathrm{KL}(q\|p_q) = \mathbb{E}_{\theta\sim q}\left[\log \frac{q}{p_q}(\theta)\right] = \mathbb{E}_{\theta\sim q}\left[\log q(\theta) - \log q_{g_q}(\theta)\right] + \log \int q_{g_q}(\theta)d\theta.
\tag{32}
$$

Combining (31) and (32), we get for any $q \in \mathcal{P}$,

$$
\begin{aligned}
\mathcal{D}(g_q) &= -\frac{1}{n}\sum_{i=1}^n \ell_i^*(g_{q,i}) - \lambda \log \int q_{g_q}(\theta)d\theta \\
&= \frac{1}{n}\sum_{i=1}^n (\ell_i(h_q(x_i)) - h_q(x_i)g_{q,i}) - \lambda \mathrm{KL}(q\|p_q) + \lambda \mathbb{E}_{\theta\sim q}\left[\log q(\theta) - \log q_{g_q}(\theta)\right] \\
&= \frac{1}{n}\sum_{i=1}^n (\ell_i(h_q(x_i)) - h_q(x_i)g_{q,i}) - \lambda \mathrm{KL}(q\|p_q) + \lambda \mathbb{E}_{\theta\sim q}\left[\log q(\theta)\right] + \mathbb{E}_{\theta\sim q}\left[\frac{1}{n}\sum_{i=1}^n h_\theta(x_i)g_{q,i} + \lambda'\|\theta\|_2^2\right] \\
&= \mathcal{L}(q) - \lambda \mathrm{KL}(q\|p_q).
\end{aligned}
$$

This concludes the proof. $\qquad\square$

Proof of Theorem 4. From direct computation, we get for $q \in \mathcal{P}$,

$$\frac{\delta \mathcal{L}}{\delta q'}(q')|_{q'=p_q}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\partial_z \ell(h_{p_q}(x_i), y_i)h_\theta(x_i) + \lambda'\|\theta\|_2^2 + \lambda \log p_q(\theta)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\partial_z \ell(h_{p_q}(x_i), y_i) - \partial_z \ell(h_q(x_i), y_i))h_\theta(x_i) + const.$$

Hence, we have

$$\left|\int \frac{\delta \mathcal{L}}{\delta q'}(q')|_{q'=p_q}(\theta)(q - p_q)(\theta)\mathrm{d}\theta\right| = \left|\int \frac{1}{n}\sum_{i=1}^{n}(\partial_z \ell(h_{p_q}(x_i), y_i) - \partial_z \ell(h_q(x_i), y_i))h_\theta(x_i)(q - p_q)(\theta)\mathrm{d}\theta\right|$$

$$= \left|\frac{1}{n}\sum_{i=1}^{n}(\partial_z \ell(h_{p_q}(x_i), y_i) - \partial_z \ell(h_q(x_i), y_i))(h_q(x_i) - h_{p_q}(x_i))\right|$$

$$\leq \frac{C_2}{n}\sum_{i=1}^{n}|h_q(x_i) - h_{p_q}(x_i)|^2$$

$$\leq B^2 C_2 \|q - p_q\|_{L_1(\mathrm{d}\theta)}^2$$

$$\leq 2B^2 C_2 \mathrm{KL}(q\|p_q),$$

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

By the convexity of $\mathcal{L}$, we get

$$\mathcal{L}(p_q) - \mathcal{D}(g_q) - \lambda \mathrm{KL}(q\|p_q) = \mathcal{L}(p_q) - \mathcal{L}(q)$$

$$\leq -\int \frac{\delta \mathcal{L}}{\delta q'}(q')|_{q'=p_q}(\theta)(q - p_q)(\theta)\mathrm{d}\theta$$

$$\leq 2B^2 C_2 \mathrm{KL}(q\|p_q).$$

Therefore, we obtain $\mathcal{L}(p_q) - \mathcal{D}(g_q) \leq (\lambda + 2B^2 C_2)\mathrm{KL}(q\|p_q)$. $\qquad\square$

B ADDITIONAL DETAILS

B. 1 Computation of the Dual Objective

We briefly outline the estimation of the dual objective (25), which consists of the sum of Fenchel conjugate functions and the normalization term as follows:

$$\mathcal{D}(g) = -\frac{1}{n}\sum_{i=1}^{n}\ell_i^*(g_i) - \lambda \log \int q_g(\theta)\mathrm{d}\theta,$$

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

The Fenchel conjugate ℓ_i^* can be explicitly described for typical loss functions. For instance, for the squared loss function $\ell_i(z) = 0.5(z - y_i)^2$, its Fenchel conjugate is $\ell_i^*(g) = 0.5g^2 + gy_i$.

Direct computation of the normalization term of q_g is difficult in general, but it can be efficiently estimated by the following procedure. First, we reformulate this term as follows:

$$\int q_g(\theta)\mathrm{d}\theta = \int \exp\left(-\frac{1}{\lambda}\left(\frac{1}{n}\sum_{i=1}^{n}h_\theta(x_i)g_i + \lambda'\|\theta\|_2^2\right)\right)\mathrm{d}\theta$$

$$= Z\int \exp\left(-\frac{1}{\lambda n}\sum_{i=1}^{n}h_\theta(x_i)g_i\right)\frac{\exp\left(-\frac{\lambda'}{\lambda}\|\theta\|_2^2\right)}{Z}\mathrm{d}\theta,$$

where Z is the normalization term of the Gaussian distribution appearing above, that is, $Z = \int \exp\left(-\frac{\lambda'}{\lambda}\|\theta\|_2^2\right) \mathrm{d}\theta$. Hence, we can estimate the normalization term of q_g by the expectation with respect to the Gaussian distribution and computation of Z. This expectation can be approximated by obtaining samples from the Gaussian distribution in proportion to $\exp\left(-\frac{\lambda'}{\lambda}\|\theta\|_2^2\right)$, and its normalization term is exactly $Z = \left(\frac{\pi\lambda}{\lambda'}\right)^{d/2}$. We can use the same samples from the Gaussian distribution for computing this expectation to reduce the computational cost. Note that the normalization constant of the proximal Gibbs distribution p_q can be approximated via the same procedure.