CONVERGENCE OF THE CAVI ALGORITHM FOR LOG-CONCAVE MEASURES

Dwaipayan Saha

Abstract. — We survey [**AL24**] which discusses the method of mean field variational inference (MFVI), a technique utilized to approximate a high-dimensional probability distribution using a product of simpler, factorized measures. This is done to minimize the relative entropy with respect to a complex distribution, denoted as ρ. The Coordinate Ascent Variational Inference (CAVI) is a technique that incrementally optimizes this approximation by focusing on one factor at a time. The authors, Arnese and Lacker, detail convergence results of the CAVI algorithm which were not well-understood prior. Specifically, they explore reference measures ρ that admit a log-concave density and detail differing convergence results based off various smoothness conditions placed on $\log \rho(dx)$. Interestingly, despite the MFVI problem being a non-convex optimization problem, when ρ is a log-concave measure, it is "displacement convex" in the sense of optimal transport, allowing for convergence proofs of the same flavor as the infamous descent lemma in the Euclidean setting.

Contents

1. Introduction and Preliminaries

1.1. Overview. — Variational inference (VI), is a computational technique used extensively in Bayesian statistics as an alternative to Markov Chain Monte Carlo methods. The primary objective of VI is to find a simpler, factorized probability measure (product measure) from a specified family that minimizes relative entropy, or Kullback-Leibler divergence, to a given complex measure ρ. This is expressed mathematically as finding the infimum of the relative entropy between the approximating and target measures.

Different families of measures used in VI, like Gaussian measures or, in our case, product measures, necessitate specific algorithms tailored for optimization. Among these, the Coordinate Ascent Variational Inference (CAVI) algorithm is highlighted, which optimizes

Key words and phrases. — CAVI, Relative Entropy, Mean Field Family.

a selected marginal distribution while holding others constant. In fact, the algorithm we will analyze can be viewed as a Block Coordinate Descent algorithm in the Wasserstein Space.

In this paper, we survey results that explore the convergence behavior of the CAVI algorithm, particularly for log-concave density $\rho(dx) = e^{-\psi(x)}dx$, where it has been observed to have favorable convergence properties. We mainly focus on recent work by Arnese and Lacker [**AL24**] who argue convergence to a minimizer under some integrability conditions on ψ. Furthermore, they show linear convergence rate if ψ has Lipschitz gradient and exponential convergence rate when ψ is strongly convex. The analysis follows by treating the mean field VI problem as a geodesically convex problem in the Wasserstein space (a space used to measure distances between probability distributions), which is effective when the target density is log-concave.

Overall, the work by Arnese and Lacker aligns itself within a broader research context that connects optimization techniques with the geometry of Wasserstein space, often employed in sampling and gradient flows in statistical learning [**San16, Lac23**]. In showing the linear and exponential convergence rates, they employ Wasserstein gradients using Otto Calculus in place of Euclidean gradients. We omit such proofs in this survey for brevity and direct the interested reader to [**AL24**, Section 5, 6].

1.2. Model and Optimization Problem. — The target reference probability measure ρ on $\mathbb{R}^k$ is assumed to be log-concave, i.e., it has density function $e^{-\psi(x)}$, where $\psi : \mathbb{R}^k \to \mathbb{R}$ is a convex function. [1] We partition our variables $1, \dots, k$ into d blocks, and assume that each block has dimension k_i with $\sum_{i=1}^{k} k_i = k$. As such, any $x \in \mathbb{R}^k$ can be written as $x = (x^1, \dots, x^d)$, with $x^i \in \mathbb{R}^{k_i}$. Let $\mathcal{P}(\mathbb{R}^k)$ denote the space of probability measures on $\mathbb{R}^k$, and let $\mathcal{P}^{\otimes d}(\mathbb{R}^k)$ denote the subset of product measures, under which the blocks are independent:

$$\mathcal{P}^{\otimes d}(\mathbb{R}^k) = \{\mu^1 \otimes \cdots \otimes \mu^d : \mu^i \in \mathcal{P}(\mathbb{R}^k), i = 1, \dots, d\} \subset \mathcal{P}(\mathbb{R}^k).$$

We will routinely identify $\mathcal{P}^{\otimes d}(\mathbb{R}^k)$ with $\mathcal{P}(\mathbb{R}^{k_1}) \times \cdots \times \mathcal{P}(\mathbb{R}^{k_d})$ in the natural way, by identifying a product measure $\mu^1 \otimes \ldots \otimes \mu^d$ with the vector $(\mu^1, \dots, \mu^d)$ of marginal measures.

Definition 1.1 **(Relative Entropy).** — Let μ, ν be two probability measures over the measurable space $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$, then the relative entropy is

$$H(\mu \,||\, \nu) := \begin{cases} \int_{\mathbb{R}^k} \frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu} d\nu & \text{if } \mu \ll \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

Using our notation, we can write the MFVI optimization problem as

$$\inf_{\mu \in \mathcal{P}^{\otimes d}(\mathbb{R}^k)} H(\mu \,||\, \rho). \tag{1}$$

Definition 1.2 **(Differential Entropy).** — Let μ be a probability measure over the measurable space $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ with density function $\frac{d\mu}{d\lambda}$, then the (negative) differential

1. As in, for any $x, y \in \mathbb{R}^k, t \in [0, 1]$, we have $\psi((1-t)x + ty) \leq (1-t)\psi(x) + t\psi(y)$. Sometimes we assume ψ to be λ-strongly convex, which means that the map $x \mapsto \psi(x) - \frac{\lambda}{2}\|x\|^2$ is convex.

entropy is [2]

$$h(\mu) := \begin{cases} \int_{\mathbb{R}^k} \frac{d\mu}{d\lambda} \log \frac{d\mu}{d\lambda} d\lambda & \text{if } \mu \text{ admits a density with } \frac{d\mu}{d\lambda} \log \frac{d\mu}{d\lambda} \in L^1(\mathbb{R}^k) \\ +\infty & \text{otherwise.} \end{cases}$$

Notice that we defined h above be the negative differential entropy. This is because for product measures $\mu = \mu^1 \otimes \cdots \otimes \mu^d$, we know that h tensorizes as

$$h(\mu) = \sum_{i=1}^{n} h(\mu^i).$$

Proposition 1.3. — *For any product measure $\mu = \mu^1 \otimes \cdots \otimes \mu^d$ with density $\frac{d\mu}{d\lambda}$ and log-concave measure ρ with density function $e^{-\psi(x)}$ on $\mathbb{R}^k$, we have*

$$H(\mu \,||\, \rho) = \sum_{i=1}^{d} h(\mu^i) + \int_{\mathbb{R}^k} \psi(x)\mu(dx).$$

Proof. — This follows from the definition of relative entropy:

$$\begin{aligned} H(\mu \,||\, \rho) &= \int_{\mathbb{R}^k} \log \frac{d\mu}{d\rho} \mu(dx) = \int_{\mathbb{R}^k} \log \frac{d\mu}{d\lambda} \mu(dx) - \int_{\mathbb{R}^k} \log \frac{d\rho}{d\lambda} \mu(dx) \\ &= h(\mu) + \int_{\mathbb{R}^k} \psi(x)\mu(dx) \\ &= \sum_{i=1}^{d} h(\mu^i) + \int_{\mathbb{R}^k} \psi(x)\mu(dx) \end{aligned}$$

where the second equality holds due to the existence of the required densities, and the fourth by tensorization of negative differential entropy. □

1.3. Wasserstein Distance and Optimal Transport. — This section serves as quick recall on definitions and required results for the remainder of the paper. Denote $\mathcal{P}_2(\mathbb{R}^k)$ to be the set of probability measures on $\mathbb{R}^k$ with finite second moment and $\mathcal{P}_2^{\otimes d}(\mathbb{R}^k) = \mathcal{P}_2(\mathbb{R}^k) \cap \mathcal{P}^{\otimes d}(\mathbb{R}^k)$. Moreover, denote by $\mathcal{P}_{2,ac}(\mathbb{R}^k)$ the set of measures with finite second moment that admit a density with respect to the Lebesgue measure on $\mathbb{R}^k$, and $\mathcal{P}_2^{\otimes d}(\mathbb{R}^k) = \mathcal{P}_2^{\otimes d}(\mathbb{R}^k) \cap \mathcal{P}_{2,ac}(\mathbb{R}^k)$. Similarly, we write $\mathcal{P}_p(\mathbb{R}^k)$ for the set of probability measures on $\mathbb{R}^k$ with finite pth moment, and we let $\mathcal{P}_p^{\otimes d}(\mathbb{R}^k) = \mathcal{P}_p^{\otimes d}(\mathbb{R}^k) \cap \mathcal{P}_p(\mathbb{R}^k)$.

Definition 1.4 **(Coupling).** — Let μ, ν be two probability measures, and let

$$\mathcal{C}(\mu, \nu) := \{\text{Law}(X, Y) : X \sim \mu, Y \sim \nu\}.$$

Any probability measure $\pi \in \mathcal{C}(\mu, \nu)$ is called a *coupling* of μ, ν.

The notion of couplings allows us to define the quadratic Wasserstein distance. We define the quadratic Wasserstein distance on $\mathcal{P}_2(\mathbb{R}^k)$ by

$$\mathbb{W}_2(\mu, \nu) = \inf_{\pi \in \mathcal{C}(\mu,\nu)} \left(\int_{\mathbb{R}^k \times \mathbb{R}^k} \|x - y\|^2 \, \pi(dx, dy) \right)^{1/2}.$$

2. This means that for any $A \in \mathcal{B}(\mathbb{R}^k)$, we have $\mu(A) = \int_A \frac{d\mu}{d\lambda} d\lambda$ where λ is Lebesgue measure on $\mathbb{R}^k$. Notice that this requires $\mu \ll \lambda$, which follows since differential entropy is only defined for laws of continuous random variables.

Alternatively,

$$\mathbb{W}_2(\mu,\nu) = \inf_{\pi\in\mathcal{C}(\mu,\nu)} \sqrt{\mathbf{E}_\pi[\|X-Y\|^2]},$$

where the couple $(X,Y)\sim\pi$ and $X\sim\mu, Y\sim\nu$. Furthermore, if $\mu\in\mathcal{P}_{2,ac}(\mathbb{R}^k)$, then there exists a unique μ-almost everywhere measurable function $T_{\mu\to\nu}:\mathbb{R}^k\to\mathbb{R}^k$, such that the pushforward of μ by $T_{\nu\to\mu}$ is ν and

$$\mathbb{W}_2^2(\mu,\nu) = \int_{\mathbb{R}^k} \|x - T_{\mu\to\nu}(x)\|^2 \,\mu(dx).$$

This map $T_{\mu\to\nu}$ is known as the Brenier map and can also be characterized as the unique gradient of a convex function which pushes forward μ to ν [**Vil08**, Theorem 2.12].

1.4. Geodesic Convexity and Subdifferential Calculus. — Given two measures $\mu_0,\mu_1\in\mathcal{P}_{2,ac}(\mathbb{R}^k)$ and a variable $t\in(0,1)$, define the pushforward of μ_0 to μ_1 by the map $x\mapsto(1-t)x + tT_{\mu_0\to\mu_1}(x)$. The resulting path $(\mu_t)_{t\in[0,1]}$ forms the unique geodesic connecting μ_0 to μ_1, which can be thought of as a generalization of lines on manifolds. A subset $S\subseteq\mathcal{P}_{2,ac}(\mathbb{R}^k)$ is said to be geodesically convex if every geodesic $(\mu_t)_{t\in[0,1]}$ connecting any μ_0,μ_1 in S remains entirely within S. Furthermore, a functional $\Phi:\mathcal{P}_{2,ac}(\mathbb{R}^k)\to\mathbb{R}$ is considered λ-geodesically convex if, for any $\mu_0,\mu_1\in\mathcal{P}_{2,ac}(\mathbb{R}^k)$ and any $t\in[0,1]$, the following inequality holds: [3]

$$\Phi(\mu_t) \le (1-t)\Phi(\mu_0) + t\Phi(\mu_1) - \frac{t(1-t)\lambda}{2}\mathbb{W}_2^2(\mu_1,\mu_0).$$

It is known that $\mathcal{P}_2^{\otimes d}(\mathbb{R}^k)$ is geodesically convex set [**Lac23**], which suits our needs since our iterates have finite pth moments. Furthermore, it is a well known fact that ρ is log-concave if and only if $H(\cdot\,||\,\rho)$ is a geodesically convex functional over $\mathcal{P}_2^{\otimes d}(\mathbb{R}^k)$ [**AGS**]. These two facts allow us to understand why Problem (1) being a non-convex optimization problem, in the usual sense, is a convex optimization problem since it minimizes a geodesically convex function over a geodesically convex set.

Some proofs presented in [**AL24**] require results from subdifferential calculus in Wasserstein space. They state various theorems that establish conditions on which the functional h and $H(\cdot\,||\,\rho)$ have non-empty subdifferential and the forms of their respective Wasserstein Gradients from [**AGS**].

1.5. Main Theorem. —

Theorem 1.5. — *Let $\psi:\mathbb{R}^k\to\mathbb{R}$ be a convex function such that $\rho(x)\propto e^{-\psi(x)}$ defines a probability density, and assume there exist finite constants $c>0$ and $p\ge 2$ such that*

$$\|\psi(x)\| \le c(1+\|x\|^p), \quad \textit{and} \quad \|\nabla\psi(x)\| \le c(1+\|x\|^p), \quad \textit{for almost every } x\in\mathbb{R}^k,$$

where $\nabla\psi$ is the weak gradient. Let $\mu_0\in\mathcal{P}^{\otimes d}(\mathbb{R}^{k_i})$ have finite pth moment, and define the CAVI iterates $\mu_n=\bigotimes_{i=1}^d\mu_n^i$ as in (1.4). Then $H(\mu_1\,||\,\rho)<\infty$, and the following hold:

1. *The sequence (μ_n) is tight, and every weak limit point is a minimizer of (1).*
2. *If ψ is strictly convex, then (1) admits a unique minimizer μ_*, and $\mu_n\to\mu_*$ weakly.*

3. Note that $\lambda=0$ implies that Φ is geodesically convex and λ-strictly convex if the above inequality is strict.

3. *If ψ is differentiable and $\nabla\psi$ is L-Lipschitz for $L > 0$, and if μ_* is a minimizer of* (1.3)*, then*

$$H(\mu_n \,||\, \rho) - H(\mu_* \,||\, \rho) \leq \left(2 + H(\mu_1 \,||\, \rho) - H(\mu_* \,||\, \rho) + \frac{1}{R\sqrt{Ld}}\right) \cdot \frac{2R^2Ld}{n},$$

where $R := \sup_{n\in\mathbb{N}} \mathbb{W}_2(\mu_n, \mu_*) < \infty$.

4. *If ψ is λ-strongly convex with L-Lipschitz gradient, for $L \geq \lambda > 0$, then* (1.3) *admits a unique minimizer μ_*, and*

$$\frac{\lambda}{2}\mathbb{W}_2^2(\mu_n, \mu_*) \leq H(\mu_n \,||\, \rho) - H(\mu_* \,||\, \rho) \leq \left(1 - \frac{\lambda^2}{L^2d + \lambda^2}\right)^{n-1} (H(\mu_1 \,||\, \rho) - H(\mu_* \,||\, \rho)).$$

We do not prove all four components of this Theorem in this survey. Rather, we prove the first two parts. The third and fourth result regarding the rate of convergence rely on Wasserstein Calculus and Geodesic Convexity, and so we state them without proof. However, the proofs are quite similar to their Euclidean descent lemma counterpart. Lastly, notice for the linear convergence rate, we have an R which represents the maximum quadratic Wasserstein distance between any of the iterates (μ_n) and the reference measure ρ. The result would be useless if this "diameter" was not finite. It turns out when ρ is strongly log-concave or satisfies other conditions (that we shall detail later), we can prove R to be finite using tools from concentration of measure.

2. CAVI Algorithm

The CAVI algorithm is an iterative algorithm intending to solve the Optimization Problem (1). We will impose some conditions on the convex measurable function $\psi : \mathbb{R}^k \to \mathbb{R}$:[4]

(Q.1) We can find a $a \in \mathbb{R}$ and $\beta > 0$ such that $\psi(x) \geq a + \beta\|x\|$ almost everywhere.

(Q.2) There exists $c > 0$ and $p \geq 1$ such that $\psi(x) \leq c(1 + \|x\|^p)$ almost everywhere.

(Q.3) $\int_{\mathbb{R}^k} e^{-\psi(x)}\, dx = 1$, so that $\rho(dx) = e^{-\psi(x)}dx$ defines a probability density function.

It is easy to check for any log-concave measure $\rho = e^{-\psi}$ that Q.1 holds since ψ is convex. Now, we initialize $\mu_0 = \bigotimes_{i=1}^d \mu_0^i$ to be any measure in $\mathcal{P}^{\otimes d}(\mathbb{R}^k)$. The update of coordinate $i \in [d]$ at iteration n is as follows:

$$\mu_{n+1}^i = \underset{\nu\in\mathcal{P}(\mathbb{R}^{k_i})}{\arg\min}\, H(\mu_{n+1}^1 \otimes \ldots \otimes \mu_{n+1}^{i-1} \otimes \nu \otimes \mu_n^{i+1} \otimes \ldots \otimes \mu_n^d \,||\, p),$$

the solution to which is a sub-exponential distribution that will be detailed later. We will use the following notation throughout the remainder of the paper for brevity. To that end, for any $x = (x^1, \ldots, x^d) \in \mathbb{R}^k = \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_d}, i \in [d]$, and $y \in \mathbb{R}^{k_i}$, we denote

$$x^{-i} = (x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^d) \quad \text{and} \quad (y, x^{-i}) = (x^1, \ldots, x^{i-1}, y, x^{i+1}, \ldots, x^d).$$

As such, for any $i \in [d]$, we denote by $\mu^{-i} \in \mathcal{P}(\mathbb{R}^{k-k_i})$ the marginal of x^{-i}

$$\mu^{-i} = \mu^1 \otimes \cdots \otimes \mu^{i-1} \otimes \mu^{i+1} \otimes \cdots \otimes \mu^d,$$

and for any $\nu \in \mathcal{P}(\mathbb{R}^{k_i})$, we write

$$\nu \otimes \mu^{-i} = \mu^1 \otimes \cdots \otimes \mu^{i-1} \otimes \nu \otimes \mu^{i+1} \otimes \cdots \otimes \mu^d.$$

4. Note that these conditions are more general than those listed in Theorem 1.5.

Lastly, we utilize the following notation to specify intermediate measures amongst the iterates $(\mu_n)_{n\geq 0}$ of the CAVI algorithm

$$\mu_{n:i} = \mu_{n+1}^1 \otimes \cdots \otimes \mu_{n+1}^i \otimes \mu_n^{i+1} \otimes \cdots \otimes \mu_n^d,$$

where we note that that $\mu_{n:0} = \mu_n$, and $\mu_{n:d} = \mu_{n+1}$. Similar to the original paper, we use the shorthand $H(\mu \,||\, \rho) = H(\mu)$ throughout the rest of this paper.

***Lemma 2.1* ([AL24]).** — *Assume (Q.1-3). If $\mu_0 \in \mathcal{P}_p^{\otimes d}(\mathbb{R}^k)$, then the following hold for each $n \geq 1$:*[5]

1. *The iterates μ_n are well defined subexponential distributions.*
2. *$H(\mu_n) < \infty$.*
3. *For each $i \in [d]$, $\mu_{n+1}^i(x^i) \propto \exp\left(-\int_{\mathbb{R}^{k-k_i}} \psi(x^i, y^{-i})\mu_{n:i-1}^{-i}(dy^{-i})\right)$ is the unique minimizer, i.e.,*
$$\mu_{n+1}^i = \operatorname*{arg\,min}_{\eta \in \mathcal{P}(\mathbb{R}^{k_i})} H(\eta \otimes \mu_{n:i-1}^{-i}).$$
4. *$H(\mu_n) \leq H(\mu_{n-1})$ and more generally $H(\mu_{n:i}) \leq H(\mu_{n:i-1})$ for all $i \in [d]$.*

Proof. — The proofs of 1 and 2 follow from the assumptions Q.1 and Q.2. We provide a formal proof of 3 and 4. To that end, we prove 3 for $n = i = 1$ as in

$$\mu_2^1 = \operatorname*{arg\,min}_{\eta \in \mathcal{P}(\mathbb{R}^{k_i})} H(\eta \otimes \mu_1^{-1}).$$

Now, we define $f(x^1) = \int_{\mathbb{R}^{k-k_1}} \psi(x^1, y^{-1})\mu_1^{-1}(dy^{-1})$ and notice that

$$\mu_2^1(dx^1) = \frac{1}{Z} e^{-f(x_1)} dx^1,$$

for some appropriate normalization constant Z. Next, note that by Proposition 1.3

$$\begin{aligned}
H(\eta \otimes \mu_1^{-1}) &= h(\eta) + \sum_{i=2}^d h(\mu^i) + \int_{\mathbb{R}^{k_1}} \int_{\mathbb{R}^{k-k_1}} \psi(x^1, x^{-1})\mu_1^{-1}(dx^{-1})\eta(dx^1) \\
&= h(\eta) + \sum_{i=2}^d h(\mu^i) + \int_{\mathbb{R}^{k_1}} f(x^1)\eta(dx^1) \\
&= \sum_{i=2}^d h(\mu^i) + \int_{\mathbb{R}^{k_1}} \log(\eta(x^1))\eta(dx^1) - \int_{\mathbb{R}^{k_1}} \log(\mu_2^1(x^1))\eta(dx^1) - \log Z \\
&= \sum_{i=2}^d h(\mu^i) + H(\eta \,||\, \mu_2^1) - \log Z.
\end{aligned}$$

To finish, note that taking the minimum $\eta \in \mathcal{P}(\mathbb{R}^{k_i})$ on both sides effectively reduces to minimizing $H(\eta \,||\, \mu_2^1)$. This expression is uniquely minimized by $\eta = \mu_2^1$ since the relative

5. From here onward, we will often use the notation $\mu(x)$ which seems nonsensical to evaluate the measure at some point. However, note that this is, for brevity, an abuse of notation used to represent the Radon-Nikodym derivative $\frac{d\mu}{d\lambda}(x)$ or the density function corresponding to the measure μ. Additionally, we always integrate against $\mu(dx)$ which implies that μ must admit a density w.r.t to the Lebesgue measure, but this is not necessary and everything remains well defined otherwise. This is especially relevant since μ_0 can in fact be a Dirac.

entropy is lower-bounded at 0 where equality is achieved only if both measures are identical. Lastly, notice that 4 follows immediately from 3.

□

***Lemma 2.2* (Uniform Moment Bound).** — *Assume (Q.1-3). For each $q \geq 1$ and $n \in \mathbb{N}$,*

$$\int_{\mathbb{R}^k} \|x\|^q \mu_n(dx) \leq e^{(2d+1)H(\mu_1)-(d+1)\alpha} \left(\int_{\mathbb{R}^k} e^{-(\beta\|x\|+\alpha)/2}\, dx \right)^{2d} \int_{\mathbb{R}^k} \|x\|^q e^{-\beta \sum_i \|x_i\|}\, dx < \infty.$$

Proof. — To begin, we notate

$$\mathcal{S} := \{\mu \in \mathcal{P}_2^{\otimes d}(\mathbb{R}^k) : H(\mu) \leq H(\mu_1)\}.$$

Due to Lemma 2.1, we know that $\mu_{n:i} \in \mathcal{S}$ for all $n \geq 1$ and $i \in [d] \cup \{0\}$. Furthermore, by the Gibbs Variational Principle, Theorem A.4, where we choose the measure ρ and function $\psi(x)/2$, we can write for any $\mu \in \mathcal{S}$

$$\begin{aligned}
\int_{\mathbb{R}^k} \psi(x)\mu(dx) &\leq 2H(\mu) + 2\log \int_{\mathbb{R}^k} e^{\psi(x)/2}\rho(dx) \\
&\leq 2H(\mu_1) + 2\log \int_{\mathbb{R}^k} e^{-\psi(x)/2} dx \\
&\leq 2H(\mu_1) + 2\log \int_{\mathbb{R}^k} e^{-(\alpha+\beta\|x\|)/2} dx := I < \infty
\end{aligned}$$

Using Proposition 1.3,

$$\begin{aligned}
H(\mu_n) &= \int_{\mathbb{R}^k} \psi(x)\mu_n(dx) + \sum_{i=1}^d \int_{\mathbb{R}^{k_1}} \log\left(\frac{\mu_n^i(x^i)}{Z_n^i}\right) \mu_n^i(dx^i) \\
&= \int_{\mathbb{R}^k} \psi(x)\mu_n(dx) - \log Z_n - \sum_{i=1}^d \int_{\mathbb{R}^{k_1}} \int_{\mathbb{R}^{k-k_i}} \psi(x^i, y^{-i}) \mu_{n-1:i-1}^{-i}(dy^{-i}) \mu_n^i(dx^i) \\
&= \int_{\mathbb{R}^k} \psi(x)\mu_n(dx) - \log Z_n - \sum_{i=1}^d \int_{\mathbb{R}^k} \psi(x)\mu_{n-1:i}(dx) \\
&\geq \alpha - \log Z_n - dI
\end{aligned}$$

where Z_n is the normalization constant of μ_n and the Z_n^i are the normalization constants of μ_n^i for all $i \in [d]$. The last inequality holds by using Q.1 for $\psi \geq \alpha$ and the earlier bound for $\mu_{n-1:i}$ which is indeed in $\mathcal{S}$ for each $i \in [d]$. As a result, $H(\mu_1) \geq H(\mu_n) \geq \alpha - \log Z_n - dI$ which implies that $Z_n \geq e^{\alpha - dI - H(\mu_1)} > 0$. Next, note

$$\mu_n(dx) = \frac{1}{Z_n} e^{-\sum_{i=1}^d f_i(x^i)} dx, \quad \text{where} \quad f_i(x^i) = \int_{\mathbb{R}^{k-k_i}} \psi(x^i, y^{-i}) \mu_{n-1:i-1}^{-i}(dy^{-i}),$$

and that by Q.1 $\sum_{i=1}^d f_i(x^i) \geq d\alpha + \beta \sum_{i=1}^d \|x^i\|$. Thus, we have

$$\int_{\mathbb{R}^k} \|x\|^q \mu_n(dx) = \frac{1}{Z_n} \int_{\mathbb{R}^k} \|x\|^q e^{-\sum_{i=1}^d f_i(x^i)} dx \leq e^{dI + H(\mu_1) - (d+1)\alpha} \int_{\mathbb{R}^k} \|x\|^q e^{-\beta \sum_{i=1}^d \|x^i\|} dx,$$

where substituting I gives the required result.

□

Lemma 2.3. — *The sequence (μ_n) is tight. If a subsequence (μ_{n_t}) converges weakly to some μ_*, then*

$$\int_{\mathbb{R}^k} \psi(x)\mu_{n_t}(dx) \to \int_{\mathbb{R}^k} \psi(x)\mu_*(dx).$$

Proof. — By Lemma 2.1, we know that $\sup_{n\in\mathbb{N}} H(\mu_n) = H(\mu_1) < \infty$. Furthermore, we also know that the sub-level sets of relative entropy are compact [**BD19**, Lemma 2.4(C)]. This tells us that $\mathcal{S}$ defined in the previous proof is a compact subset of $\mathcal{P}_2^{\otimes d}(\mathbb{R}^k)$. In turn, let $\mu_{n_t} \to \mu_*$.

Notate (X_n) to be a sequence of random variables with laws (μ_n) and notice that $(\psi(X_n))$ is uniformly integrable sequence since

$$\begin{aligned}\lim_{M\to\infty} \sup_n \mathbf{E}[|\psi(X_n)|\mathbf{1}_{|\psi(X_n)|\geq M}] &= \lim_{M\to\infty} \sup_n \mathbf{E}\left[\frac{|\psi(X_n)|^{1+\delta}}{|\psi(X_n)|^{\delta}}\mathbf{1}_{|\psi(X_n)|\geq M}\right] \\ &\leq \lim_{M\to\infty} \sup_n \frac{1}{M^\delta}\mathbf{E}[|\psi(X_n)|^{1+\delta}] = \lim_{M\to\infty} \frac{C}{M^\delta} = 0\end{aligned}$$

where we use Lemma 2.2 to get the second to last equality. As a result, for our subsequence (μ_{n_t}) that converges weakly or in distribution to μ_*, note that

$$\begin{aligned}\mathbf{E}[\psi(X_{n_t}) - \psi(X_*)] = \mathbf{E}[\psi_M(X_{n_t}) - \psi_M(X_*)] + \mathbf{E}[(\psi(X_{n_t}) - \psi_M(X_{n_t}))\mathbf{1}_{|X_{n_t}|>M}] \\ + \mathbf{E}[(\psi_M(X_*) - \psi(X_*))\mathbf{1}_{|X_*|>M}]\end{aligned}$$

where $\psi_M(x) = \psi(x)$ if $\psi(x) \in [-M, M]$, M if $\psi(x) > M$, and $-M$ if $\psi(x) < -M$. Notice that this is a bounded and continuous function. As such, taking a limit, we see that the first term is zero due to weak convergence coupled with Portmanteau Theorem. Moreover, the latter terms go to zero as well since they are bounded by $\mathbf{E}[|\psi(X_{n_t})|\mathbf{1}_{|\psi(X_{n_t})|>M}]$ and $\mathbf{E}[|\psi(X_*)|\mathbf{1}_{|\psi(X_*)|>M}]$ respectively using Jensen's inequality. To finish, notice that both terms go to 0 due to uniform integrability giving the desired result. □

We aim to explore the convergence of the iterates (μ_n) (or a subsequence) and determine if such a limit point is a minimizer. For general ρ, this is not possible and, even when ρ is a log-concave measure, this is uncertain. Nevertheless, we can first establish a relationship between such a limit point and the concept of a stationary point of Problem 1, which we introduce next.

Definition 2.4 **(Stationary point).** — A stationary point $\mu \in \mathcal{P}_p^{\otimes d}(\mathbb{R}^k)$ to Problem 1 is such that $H(\mu) < \infty$ and for any $\nu^i \in \mathcal{P}_p(\mathbb{R}^{k_i})$ we have

$$H(\nu^i \otimes \mu^{-i}) \geq H(\mu).$$

Lemma 2.5. — *Assume Q.1-3. The limit points of the tight sequence (μ_n) are stationary points of H.*

Proof. — Note that Lemma 2.3 gives that (μ_n) is tight. Thus, we let (μ_{n_t}) be a subsequence converging weakly to some limit point λ. The proof strategy is to show that up to extracting a subsequence, $(\mu_{n_t:1}), (\mu_{n_t:2}), \ldots, (\mu_{n_t+1})$ all converge to the same limit point λ, from which we deduce that λ must be a stationary point.

To start off, suppose $(\mu_{n_t:1})$ converges to some limit point $\sigma \neq \lambda$. Since μ_{n_t} and $\mu_{n_t:1}$ share the same ith marginals for $i \geq 2$, we know that $\sigma_i = \lambda_i$ for $i \geq 2$. Further, denote $\eta_t = \frac{1}{2}(\mu_{n_t} + \mu_{n_t+1})$. We can then express η_t as

$$\eta_t = \frac{1}{2}(\mu^1_{n_t:1} + \mu^1_{n_t}) \otimes \mu^{-1}_{n_t},$$

which shows that $\eta_t \in \mathcal{P}_p^{\otimes d}(\mathbb{R}^k)$. Now, let $\eta = \frac{1}{2}(\lambda + \sigma)$ be its limit. Since $H(\mu_{n_t})$ is non-increasing in each update and lower bounded at 0, it follows that $H(\mu_{n_t}) - H(\mu_{n_t:1}) \to 0$. Using Proposition 1.3 and noticing that μ_{n_t} and $\mu_{n_t:1}$ share the same ith marginals for $i \geq 2$ and $\mu^1_{n_t:1} = \mu^1_{n_t+1}$, we know that

$$\lim_{t\to\infty} \int_{\mathbb{R}^k} \psi(x)\mu_{n_t}(dx) + h(\mu^1_{n_t}) = \lim_{t\to\infty} \int_{\mathbb{R}^k} \psi(x)\mu_{n_t:1}(dx) + h(\mu^1_{n_t+1}) := \ell \tag{2}$$

By the convexity of H and noticing that μ_{n_t} and η share the same ith marginals for $i \geq 2$, it is clear that

$$\int_{\mathbb{R}^k} \psi(x)\eta_t(dx) + h(\eta^1_t) \leq \int_{\mathbb{R}^k} \psi(x)\mu_{n_t}(dx) + h(\mu^1_{n_t}).$$

Now, by Lemma 2.3 and lower semicontinuity of differential entropy and then applying above bound,

$$\int_{\mathbb{R}^k} \psi(x)\eta(dx) + h(\eta^1) \leq \liminf_{t\to\infty} \int_{\mathbb{R}^k} \psi(x)\eta_t(dx) + h(\eta^1_t) \leq \liminf_{t\to\infty} \int_{\mathbb{R}^k} \psi(x)\mu_{n_t}(dx) + h(\mu^1_{n_t}) = \ell,$$

where the last equality holds due to Equation 2. Suppose for the sake of contradiction that $\int_{\mathbb{R}^k} \psi(x)\eta(dx) + h(\eta^1) < \ell$ from which we get

$$\begin{aligned}\lim_{t\to\infty} \int_{\mathbb{R}^k} \psi(x)(\eta^1 \otimes \mu^{-1}_{n_t})(dx) + h(\eta^1) &= \int_{\mathbb{R}^k} \psi(x)\eta(dx) + h(\eta^1) \\ &< \ell = \lim_{t\to\infty} \int_{\mathbb{R}^k} \psi(x)\mu_{n_t:1}(dx) + h(\mu^1_{n_t+1}),\end{aligned}$$

where we used Lemma 2.3 in the first equality and Equation 2 in the last equality. Therefore, there exists large enough t_0 such that, for all $t \geq t_0$, we have

$$\int_{\mathbb{R}^k} \psi(x)(\eta^1 \otimes \mu^{-1}_{n_t})(dx) + h(\eta^1) < \int_{\mathbb{R}^k} \psi(x)\mu_{n_t:1}(dx) + h(\mu^1_{n_t+1}) \to H(\eta^1 \otimes \mu^{-1}_{n_t}) < H(\mu_{n_t:1}),$$

where the implication follows by adding back the equivalent marginals and noticing once again that $\mu^1_{n_t+1} = \mu^1_{n_t:1}$. However, this violates the non-increasing property of H giving a contradiction. Thus, $\int_{\mathbb{R}^k} \psi(x)\eta(dx) + h(\eta^1) = \ell$. Using strict convexity of h in the usual sense, we write

$$\begin{aligned}\ell = \int_{\mathbb{R}^k} \psi(x)\eta(dx) + h(\eta^1) &< \frac{1}{2}\left(\int_{\mathbb{R}^k} \psi(x)\lambda(dx) + \int_{\mathbb{R}^k} \psi(x)\sigma(dx)\right) + \frac{1}{2}\left(h(\sigma^1) + h(\lambda^1)\right) \\ &\leq \liminf_{t\to\infty} \frac{1}{2}\left(\int_{\mathbb{R}^k} \psi(x)\mu_{n_t:1}(dx) + h(\mu^1_{n_t:1})\right) \\ &\quad + \frac{1}{2}\left(\int_{\mathbb{R}^k} \psi(s)\mu_{n_t}(dx) + h(\mu^1_{n_t})\right) = \frac{\ell}{2} + \frac{\ell}{2} = \ell\end{aligned}$$

where we used Lemma 2.3 and lower semicontinuity of differential entropy for second line followed by Equation 2 in the last line, yielding a contradiction to our original assumption. Therefore, $\lambda = \sigma$. Repeating the argument for showing that (μ_{n_t}) and $(\mu_{n_t:1})$ have the

same limit up to a subsequence iteratively gives that $(\mu_{n_t:2}), \dots, (\mu_{n_t+1})$ all have the same limit λ up to a subsequence, which we exploit in the next part.

To show that such a limit point λ must be a stationary point, consider any $\eta \in \mathcal{P}_p(\mathbb{R}^{k_i})$ and $i \in [d]$. Notice that $H(\mu_{n_t:i}) \leq H(\eta \otimes \mu^{-1}_{n_t:i})$ due to the non-decreasing property. Removing the identical marginals gives

$$\int_{\mathbb{R}^k} \psi(x)\mu_{n_t:i}(dx) + h(\mu^i_{n_t:i}) \leq \int_{\mathbb{R}^k} \psi(x)(\eta \otimes \mu^{-i}_{n_t:i})(dx) + h(\eta). \tag{3}$$

To conclude by using Lemma 2.3 and lower semicontinuity of differential entropy, we find

$$\begin{aligned}\int_{\mathbb{R}^k} \psi(x)\lambda(dx) + h(\lambda^i) &\leq \liminf_{t\to\infty} \int_{\mathbb{R}^k} \psi(x)\mu_{n_t:i}(dx) + h(\mu^i_{n_t:i}) \\ &\leq \lim_{t\to\infty} \int_{\mathbb{R}^k} \psi(x)(\eta \otimes \mu^{-i}_{n_t:i})(dx) + h(\eta) = \int_{\mathbb{R}^k} \psi(x)(\eta \otimes \lambda^{-i})(dx) + h(\eta).\end{aligned}$$

However, adding back the marginals $\sum_{j\neq i} h(\lambda^j)$ gives $H(\lambda) \leq H(\eta \otimes \lambda^{-i})$ proves that λ is indeed a stationary point. □

It turns out that stationary points also have a connection to measures satisfying a mean field equation, one that looks very similar to our optimal CAVI updates.

Lemma 2.6. — *Assume Q.1-3, and let $\mu \in \mathcal{P}_p^{\otimes d}(\mathbb{R}^k)$. Then μ is a stationary point if and only if it satisfies the following mean field equation:*

$$\mu^i(x^i) \propto \exp\left\{-\int_{\mathbb{R}^{k-k_i}} \psi(x^i, y^{-i})\mu^{-i}(dy^{-i})\right\}, \quad \forall i \in [d].$$

In this case, the measure μ is subexponential.

Proof. — We prove the foward direction first. To that end assume that μ is a stationary point and consider some $i \in [d]$.

By stationarity, for any $\nu^i \in \mathcal{P}_p^{\otimes d}(\mathbb{R}^k)$, we have $H(\nu^i \otimes \mu^{-i}) \geq H(\mu)$ and $H(\mu) < \infty$. Notice that $H(\nu^i \otimes \mu^{-i}) < \infty$ if and only if $h(\nu^i) < \infty$ since we can decompose the relative entropy by Proposition 1.3, noting the integral term is always finite since by Q.2 $\psi \in L^p(\nu^i \otimes \mu^{-i})$, and $h(\mu^i) < \infty$ for all $i \in [d]$ by stationarity. Therefore, the stationarity inequality can be rewritten by using Proposition 1.3 to

$$\int_{\mathbb{R}^{k_i}} \int_{\mathbb{R}^{k-k_i}} \psi(x^i, x^{-i})\mu(dx^{-i})\mu^i(dx^i) + h(\mu^i) \leq \int_{\mathbb{R}^{k_i}} \int_{\mathbb{R}^{k-k_i}} \psi(x^i, x^{-i})\mu(dx^{-i})\nu^i(dx^i) + h(\nu^i)$$

by subtracting the equivalent differential entropy terms. We may also write this as

$$\int_{\mathbb{R}^{k_i}} f_i(x_i)\mu^i(dx^i) + h(\mu^i) \leq \int_{\mathbb{R}^{k_i}} f_i(x_i)\nu^i(dx^i) + h(\nu^i), \tag{4}$$

where

$$f_i(x^i) := \int_{\mathbb{R}^{k-k_i}} \psi(x^i, x^{-i})\mu(dx^{-i}).$$

Define $Z = \int_{\mathbb{R}^{k_i}} e^{-f_i(x^i)}dx^i$ to be a normalization constant. By adding $\log Z$ on both sides of Inequality 4 doing an exp-log trick we can write

$$\int_{\mathbb{R}^{k_i}} -\log\left(\frac{1}{Z}e^{-f_i(x^i)}\right)\mu^i(dx^i) + h(\mu^i) \leq \int_{\mathbb{R}^{k_i}} -\log\left(\frac{1}{Z}e^{-f_i(x^i)}\right)\nu^i(dx^i) + h(\nu^i). \tag{5}$$

Consider the measure η to be such that $\eta(dx^i) = \frac{1}{Z}e^{-f_i(x^i)}dx^i$, which is well-defined by Q.1-2 and satisfies the mean field equation. Using Proposition 1.3, we can see that Inequality

5 is simply $H(\mu^i \,||\, \eta) \leq H(\nu^i \,||\, \eta)$ for all $\nu^i \in \mathcal{P}_p(\mathbb{R}^{k_i})$. This is only possible when the left hand side attains its minimum, which uniquely occurs when $\mu^i = \eta$. Repeating this argument for all $i \in [d]$ gives that μ satisfies the mean-field equation.

For the reverse direction, suppose that μ satisfies the mean field equation. In turn, initializing $\mu_0 = \mu$ gives that $\mu_n = \mu$ for all $n \in \mathbb{N}$ by definition of our updates implying that $H(\mu) < \infty$. As such, we also know μ is a fixed point for the CAVI algorithm which gives us stationarity. □

Lemma 2.6 alone seems like an irrelevant fact. However, it turns out to be a helpful tool in proving the last bridge between stationary points and minimizers to Problem 1.

Lemma 2.7. — *Under the assumptions of Theorem 1.5, $\mu \in \mathcal{P}_p^{\otimes d}(\mathbb{R}^k)$ is a stationary point if and only if it is a minimizer for the Problem 1.*

Proof. — This proof will require some of the geometric notions alluded to in Section 1.4. As such, we defer the interested reader to [**AL24**, Proposition 4.4]. □

At this point, we know that Lemma 2.3 gives us that (μ_n) is a tight sequence. Lemma 2.5 tells us that for our tight sequence of iterates (μ_n), any limiting point must be stationary point. Lastly, Lemma 2.7 tells us that a point is stationary if and only if it is a minimizer to Problem (1). This proves Part (1) of Theorem 1.5. Proving Part (2) follows since strict convexity of ψ gives that H is also strictly geodesically convex [**AGS**] and, as a result, yields at most one minimizer on the geodesically convex set $\mathcal{P}_2^{\otimes d}(\mathbb{R}^k)$ [**Lac23**], a result very similar to what is expected in the Euclidean setting.

3. Lipschitz Gradient - Linear Convergence Rate

The previous section established parts (1) and (2) of Theorem 1.5, i.e., we have convergence of our algorithm to a minimizer. Now we impose the additional L-lipschitz gradient constraint to get results regarding the *rate of convergence.*

Lemma 3.1. — *For $n \geq 1$,*

$$H(\mu_n) - H(\mu_*) \leq \frac{2LR^2 d}{n},$$

where we define $R = \sup_{n\in\mathbb{N}} \mathbb{W}_2(\mu_n, \mu_*)$, as well as $a = \frac{1}{2LR^2 d}$ and

$$C = \max\left\{2, \frac{4a(H(\mu_1) - H(\mu_*))}{1 + \sqrt{1 + 4a(H(\mu_1) - H(\mu_*))}}\right\}.$$

This establishes the required linear rate of convergence. In fact, this is slightly better than Theorem 1.5 (3). The proof of Lemma 3.1 is quite involved and requires heavy machinery from Wasserstein Calculus. We defer the interested reader to [**AL24**, Section 5].

3.1. Is this convergence rate useful? — Notice that, despite having $H(\mu_n)$ be a non-increasing sequence, we cannot easily conclude if $\sup_{n\in\mathbb{N}} \mathbb{W}_2(\mu_n, \mu_*)$ is finite. This is essential as linear rate of convergence guaranteed above included this value as R and, if such $R = \infty$, then our result is useless. Nevertheless, we see that for log concave measures ρ, this quantity is indeed finite despite being exponential in d. Imposing stronger conditions

on ρ allow us to gain significantly nicer bounds by leveraging tools from high-dimensional probability.

Proposition 3.2. — *Consider any log-concave reference measure ρ and let μ_* be the minimizer to the corresponding Problem (1). Denote $R := \sup_{n\in\mathbb{N}} \mathbb{W}_2(\mu_n, \mu_*)$. We have*

$$R \leq 2e^{\frac{1}{2}(2d+1)H(\mu_1)-\frac{1}{2}(d+1)\alpha} \left(\int_{\mathbb{R}^k} e^{-(\beta\|x\|+\alpha)/2}\, dx\right)^d \sqrt{\int_{\mathbb{R}^k} \|x\|^2 e^{-\beta\sum_i \|x^i\|}\, dx}.$$

Proof. — For any $\pi \in \mathcal{C}(\mu_n, \mu_*)$, we can write for any $n \in \mathbb{N}$

$$\mathbb{W}_2(\mu_n, \mu_*) \leq \sqrt{\mathbf{E}_\pi[\|X - Y\|^2]}$$

where the couple $(X, Y) \sim \pi$ and $X \sim \mu_n, Y \sim \mu_*$. Expanding

$$\begin{aligned}
\mathbb{W}_2(\mu_n, \mu_*) &\leq \left(\int_{\mathbb{R}^k\times\mathbb{R}^k} \|x - y\|^2 \pi(dx, dy)\right)^{1/2} \\
&\leq \left(\int_{\mathbb{R}^k\times\mathbb{R}^k} (\|x\| + \|y\|)^2 \pi(dx, dy)\right)^{1/2} \\
&= \big\|\|X\| + \|Y\|\big\|_2 \\
&\leq \big\|\|X\|\big\|_2 + \big\|\|Y\|\big\|_2
\end{aligned}$$

where we used triangle inequality on the second line and Minkowski in the last line. Notice that, by Lemma 2.2,

$$\begin{aligned}
\big\|\|X\|\big\|_2 &= \left(\int_{\mathbb{R}^k} \|x\|^2 \mu_n(dx)\right)^{1/2} \\
&\leq e^{\frac{1}{2}(2d+1)H(\mu_1)-\frac{1}{2}(d+1)\alpha} \left(\int_{\mathbb{R}^k} e^{-(\beta\|x\|+\alpha)/2}\, dx\right)^d \sqrt{\int_{\mathbb{R}^k} \|x\|^2 e^{-\beta\sum_i \|x^i\|}\, dx}.
\end{aligned}$$

Since μ_* is a fixed point of the CAVI algorithm, we can bound the second term with the same quantity above giving the desired result. □

With the above result, we claim that for our purposes $R < \infty$. However, as noted earlier, despite our iterates being contained in a $\mathbb{W}_2$-ball, our best current bound on R is exponential in d. In turn, strong convexity of ψ will allow us to do better.

Proposition 3.3. — *Let ρ be a log concave measure, i.e., $\rho(dx) = e^{-\psi(x)}dx$ where ψ is λ-strongly convex for $\lambda > 0$. Then ρ satisfies a log-Sobolev Inequality, namely*

$$Ent_\rho[e^f] \leq \frac{1}{2\lambda}\mathbf{E}_\rho[\|\nabla f\|^2 e^f]$$

for all measurable $f : \mathbb{R}^k \to \mathbb{R}$.

Proof. — We are given ρ to be a probability measure on $\mathbb{R}^k$ with density $\rho(dx) = e^{-\psi(x)}dx$ where ψ is a λ-uniformly convex function. It is known that ρ is the stationary measure of the Langevin stochastic differential equation, which is a Markov process with the generator

$$\mathscr{L}f(x) = -\sum_{i=1}^n \frac{\partial\psi(x)}{\partial x_i}\frac{\partial f(x)}{\partial x_i} + \sum_{i=1}^n \frac{\partial^2 f(x)}{\partial x_i^2}$$

written compactly $\mathscr{L}f = \Delta f - \langle \nabla\psi, \nabla f\rangle$. It is known that the corresponding carré de champ is $\Gamma(f,g) = \langle \nabla f, \nabla g\rangle$ and the iterated carré de champ can be written as

$$\begin{aligned}\Gamma_2(f,f) &= \frac{1}{2}\left(\mathscr{L}\Gamma(f,f) - 2\Gamma(f,\mathscr{L}f)\right)\\ &= \frac{1}{2}\left(\mathscr{L}(\|\nabla f\|^2) - 2\langle \nabla f, \nabla(\mathscr{L}f)\rangle\right)\end{aligned}$$

We first compute

$$\begin{aligned}\frac{1}{2}\mathscr{L}(\|\nabla f\|^2) &= \frac{1}{2}\Delta(\|\nabla f\|^2) - \frac{1}{2}\langle \nabla\psi, \nabla(\|\nabla f\|^2)\rangle\\ &= \langle \nabla\Delta f, \nabla f\rangle + \|\nabla^2 f\|_F^2 - \langle \nabla\psi, \nabla^2 f \nabla f\rangle\end{aligned}$$

and then we see that

$$\begin{aligned}-\langle \nabla f, \nabla(\mathscr{L}f)\rangle &= -\langle \nabla f, \nabla(\Delta f - \langle \nabla\psi, \nabla f\rangle)\rangle\\ &= -\langle \nabla f, \nabla\Delta f\rangle + \langle \nabla f, \nabla(\langle \nabla\psi, \nabla f\rangle)\rangle\\ &= -\langle \nabla f, \nabla\Delta f\rangle + \langle \nabla f, \langle \nabla^2\psi, \nabla f\rangle + \langle \nabla^2 f \nabla\psi\rangle\rangle.\end{aligned}$$

Combining terms yields

$$\Gamma_2(f,f) = \|\nabla^2 f\|_F^2 + \langle \nabla f, \nabla^2\psi \nabla f\rangle \geq \|\nabla^2 f\|_F^2 + \lambda\|\nabla f\|^2 = \|\nabla^2 f\|_F^2 + \lambda\Gamma(f,f) \geq \lambda\Gamma(f,f)$$

where the second equality holds from λ-strong convexity, the third by definition, and the last one since the Frobenius norm is non-negative. We know that the Markov Process associated with Langevin Diffusion with the log concave stationary measure ρ satisfies the Bakry-Émery criterion with $c = 1/\lambda$. Since it is known that Langevin Markov Process is entropically ergodic, [6] we may invoke Theorem A.3 and claim that

$$\mathrm{Ent}_\rho[f] \leq \frac{1}{2\lambda}\mathcal{E}(\log f, f) = \frac{1}{2\lambda}\mathbf{E}_\rho[\nabla \log f \cdot \nabla f].$$

Plugging in e^f instead alongside chain rule gives

$$\mathrm{Ent}_\rho[e^{\alpha f}] \leq \frac{\alpha^2}{2\lambda}\mathbf{E}_\rho[\|\nabla f\|^2 e^{\alpha f}].$$

□

Proposition 3.4. — *Let ρ be a log concave measure, i.e., $\rho(dx) = e^{-\psi(x)}dx$ where ψ is λ-strongly convex for $\lambda > 0$. If ρ satisfies the log-Sobolev inequality in Proposition 3.3, then ρ satisfies the Talagrand's T_2 Inequality with constant $1/\lambda$, namely*

$$\mathbb{W}_2(\rho,\nu) \leq \sqrt{2H(\nu)/\lambda}, \quad \textit{for all } \nu.$$

Proof. — By assumption we have that

$$\mathrm{Ent}_\rho[e^{\alpha f}] \leq \frac{\alpha^2\|\|\nabla f\|^2\|_\infty}{2\lambda}\mathbf{E}_\rho[e^{\alpha f}].$$

By Herbst A.2, we know that $f(X)$ is $\|\|\nabla f\|^2\|_\infty/\lambda$-subgaussian for $X \sim \rho$. Thus, we know for all 1-Lipschitz f that $f(X)$ is $1/\lambda$-subgaussian where we use the Lipschitz property to bound the gradient norm. As a result, for $X \sim \rho$ and any 1-Lipschitz function, we know that

$$\mathbf{P}[f(X) - \mathbf{E}f(X) \geq t] \leq e^{-\frac{\lambda t^2}{2}}.$$

6. This is implied by ergodicity which is satisfied as $\mathcal{E}(f,f) = \mathbf{E}_\rho[\|\nabla f\|^2]$ which is clearly 0 if and only if f is a constant function.

Applying Gozlan A.5 with $\mathbb{X} = \mathbb{R}^k$ and using the implication from (3) with $n = 1$ to (1), we find that ρ satisfies Talagrand's T_2 inequality with constant $1/\lambda$.

□

Given these tools, we can now provide a much tighter bound on R.

Proposition 3.5. — *Consider any log-concave reference measure ρ, i.e., $\rho(dx) = e^{-\psi(x)}dx$ where ψ is λ-strongly convex for $\lambda > 0$. Let μ_* be the minimizer to the corresponding Problem 1 and denote $R := \sup_{n\in\mathbb{N}} \mathbb{W}_2(\mu_n, \mu_*)$. We have*

$$R \leq 2\sqrt{2H(\mu_1)/\lambda} < \infty.$$

Proof. — Since $\rho(dx) = e^{-\psi(dx)}dx$ requires ψ to be λ-strongly convex, we may use Proposition 3.3 and, in turn, Proposition 3.4 to claim that ρ satisfies Talagrand's T_2 inequality with constant $1/\lambda$. Thus, by triangle inequality

$$\mathbb{W}_2(\mu_n, \mu_*) \leq \mathbb{W}_2(\mu_n, \rho) + \mathbb{W}_2(\rho, \mu_*) \leq \sqrt{2H(\mu_n)/\lambda} + \sqrt{2H(\mu_*)/\lambda} \leq 2\sqrt{2H(\mu_1)/\lambda},$$

where we use that (μ_n) is non-increasing and that (μ_*) is a fixed point of the CAVI algorithm to obtain the last inequality. To finish, notice that $H(\mu_1) < \infty$ by Lemma 2.1. □

Next, we present one last approach in bounding R. Notably, we will enforce a slightly different condition on ρ rather than ψ being λ-strongly convex.

Proposition 3.6. — *Let ρ be a log concave measure, i.e., $\rho(dx) = e^{-\psi(x)}dx$ and $\lambda > 0$. Further, let μ_* be the minimizer to the corresponding Problem 1 and denote $R := \sup_{n\in\mathbb{N}} \mathbb{W}_2(\mu_n, \mu_*)$. If $\int_{\mathbb{R}^k} e^{\lambda\|x\|^2}\rho(dx) < \infty$, then*

$$R \leq \frac{2}{\sqrt{\lambda}} \left(H(\mu_1) + \log \int_{\mathbb{R}^k} e^{\lambda\|x\|^2}\rho(dx) \right)^{1/2} < \infty.$$

Proof. — In the proof of Proposition 3.2, we proved that

$$\mathbb{W}_2(\mu_n, \mu_*) \leq \left(\int_{\mathbb{R}^k} \|x\|^2 \mu_n(dx) \right)^{1/2} + \left(\int_{\mathbb{R}^k} \|y\|^2 \mu_*(dx) \right)^{1/2}.$$

We can bound each term identically using Gibbs Variational principle A.4. We bound the first term by considering measure ρ and function $\lambda\|x\|^2$, then we can write for μ_n,

$$\begin{aligned} \int_{\mathbb{R}^k} \|x\|^2 \mu_n(dx) &\leq \frac{1}{\lambda} \left(H(\mu_n) + \log \int_{\mathbb{R}^k} e^{\lambda\|x\|^2}\rho(dx) \right) \\ &\leq \frac{1}{\lambda} \left(H(\mu_1) + \log \int_{\mathbb{R}^k} e^{\lambda\|x\|^2}\rho(dx) \right). \end{aligned}$$

Bounding the other term and adding gives the desired result. □

Proposition 3.4 gives us that any log concave measure $\rho = e^{-\psi}$, where ψ is λ-strongly convex, satisfied T_2 inequality with constant $1/\lambda$. In fact, we can show that such a measure also satisfies the condition in Proposition 3.6 with constant $\lambda' \in [0, \lambda/2)$. Thus, in this context, Proposition 3.6 gives the bound

$$R \leq 2\sqrt{\left(2H(\mu_1) + 2\log \int_{\mathbb{R}^k} e^{\lambda'\|x\|^2/2}\rho(dx) \right)/\lambda'}, \quad \lambda' \in [0, \lambda/2)$$

which is worse than what we obtain by Proposition 3.4 in a trivial sense as for $\lambda' = \lambda/2$, the bound from 3.6 is not finite. In any case, Proposition 3.6 simply provides a different condition on which we can bound R.

This concludes our discussion on Parts (1-3) of Theorem 1.5. The last part establishes the exponential convergence rate, which requires strong convexity of ψ alongside the existing Lipschitz gradient condition. It turns out that this rate is independent of R. We simply state the result of Theorem 1.5 (4) as a matter of completeness and defer the interested reader to [**AL24**, Section 6] for the proof. This allows us to conclude all parts of Theorem 1.5.

Appendix A. Key Definitions and Theorems

We will list definitions and theorems that we use without proof throughout the paper. We defer the interested reader to **[vH24]**.

Theorem A.1. — *The entropy of a nonnegative random variable Z is*

$$Ent[Z] := \mathbf{E}[Z \log Z] - \mathbf{E}[Z] \log \mathbf{E}[Z].$$

***Theorem A.2* (Herbst).** — *Suppose that*

$$Ent[e^{\lambda X}] \leq \frac{\lambda^2 \sigma^2}{2} \mathbf{E}[e^{\lambda X}] \quad \textit{for all } \lambda \geq 0.$$

Then

$$\log \mathbf{E}[e^{\lambda (X - \mathbf{E}X)}] \leq \frac{\lambda^2 \sigma^2}{2} \quad \textit{for all } \lambda \geq 0.$$

***Theorem A.3* (Bakry-Émery).** — *Let P_t be an (entropically) ergodic reversible Markov semigroup that satisfies the chain rule property, and let ρ be a stationary measure. If the Bakry-Émery criterion $\Gamma(f,f) \leq c\Gamma_2(f,f)$ is satisfied, then the log-Sobolev inequality $Ent_\rho[f] \leq \frac{c}{2}\mathcal{E}(\log f, f)$ holds.*

***Theorem A.4* (Gibbs Variational Principle).** — *Given any measure μ on measurable space $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ and measurable function $f : \mathbb{R}^k \to \mathbb{R}$ we have*

$$\log \mathbf{E}_\mu[e^f] = \sup_{\nu \in \mathcal{P}(\mathbb{R}^k)} \left\{ \mathbf{E}_\nu[f] - H(\nu \,||\, \mu) \right\}.$$

***Theorem A.5* (Gozlan).** — *Let μ be a probability measure on a Polish space $(\mathbb{X}, d)$, and let $\{X_i\}$ be i.i.d. $\sim \mu$. Denote by $d_n(x,y) := \left(\sum_{i=1}^n d(x_i, y_i)^2\right)^{1/2}$ the Euclidean metric on $\mathbb{X}^n$. Then the following are equivalent:*

1. *μ satisfies the T_2-inequality on $(\mathbb{X}, d)$:*

$$\mathbb{W}_2(\mu, \nu) \leq \sqrt{2\sigma^2 H(\nu \,||\, \mu)} \quad \textit{for all } \nu.$$

2. *$\mu^{\otimes n}$ satisfies the T_1-inequality on $(\mathbb{X}^n, d_n)$ for every $n \geq 1$:*

$$W_1(\mu^{\otimes n}, \nu) \leq \sqrt{2\sigma^2 H(\nu \,||\, \mu^{\otimes n})} \quad \textit{for all } \nu, n \geq 1.$$

3. *There is a constant C such that*

$$\mathbf{P}[f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) \geq t] \leq Ce^{-t^2/2\sigma^2}$$

for every $n \geq 1$, $t \geq 0$ and 1-Lipschitz function f on $(\mathbb{X}^n, d_n)$.

References

[AGS] L. Ambrosio, N. Gigli & G. Savaré – *Gradient flows in metric spaces and in the space of probability measures*, 2. ed ed., Lectures in Mathematics ETH Zürich, Birkhäuser, OCLC: 254181287.

[AL24] M. Arnese & D. Lacker – "Convergence of coordinate ascent variational inference for log-concave measures via optimal transport", 2024.

[BD19] A. Budhiraja & P. Dupuis – *Analysis and approximation of rare events: Representations and weak convergence methods*, 01 2019.

[Lac23] D. Lacker – "Independent projections of diffusions: Gradient flows for variational inference and optimal mean field approximations", 2023.

[San16] F. Santambrogio – " Euclidean, Metric, and Wasserstein gradient flows: an overview", 2016.

[vH24] R. van Handel – "Probability in high dimension", 2024.

[Vil08] C. Villani – "Optimal transport – old and new", vol. 338, p. xxii+973, 01 2008.

June 17, 2024

Dwaipayan Saha, Department of Computer Science, Princeton University
E-mail : dsaha@princeton.edu • *Url :* https://dsaha04.github.io/