Journal of Scientific Computing (2020) 84:51
https://doi.org/10.1007/s10915-020-01303-z

Mirror Descent Algorithms for Minimizing Interacting Free Energy

Lexing Ying[1]

Received: 13 April 2020 / Revised: 18 July 2020 / Accepted: 2 September 2020

Abstract
This note considers the problem of minimizing interacting free energy. Motivated by the mirror descent algorithm, for a given interacting free energy, we propose a descent dynamics with a novel metric that takes into consideration the reference measure and the interacting term. This metric naturally suggests a monotone reparameterization of the probability measure. By discretizing the reparameterized descent dynamics with the explicit Euler method, we arrive at a new mirror-descent-type algorithm for minimizing interacting free energy. Numerical results are included to demonstrate the efficiency of the proposed algorithms.

Keywords Mirror descent algorithms · Interacting free energy · Kullback–Leibler divergence · Reverse Kullback—Leibler divergence · Hellinger divergence

1 Introduction

This paper considers the problem of minimizing free energies of the following form

$$F(p) = D(p||\mu) + \int_\Omega p(x)V(x)\mathrm{d}x + \frac{1}{2}\iint p(x)W(x,y)p(y)\mathrm{d}x\mathrm{d}y \tag{1}$$

for a probability density p over domain Ω. $D(p||\mu)$ is a divergence function between p and a reference density μ and typically examples are Kullback–Leibler divergence, reverse Kullback–Leibler divergence, and Hellinger divergence. In the interacting term $\iint pWp\mathrm{d}x\mathrm{d}y$, W is symmetric and can either be positive-definite or not. Non-positive-definite interacting terms appear in Keller–Segel models in mathematical biology and granular flows in kinetic theory. Recently, positive-definite interacting terms appear in the mean field modeling of neural network training [9,15,19,21].

The work of L.Y. is partially supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC) program and also by the National Science Foundation under award DMS-1818449. The author thanks Wuchen Li and Wotao Yin for comments and suggestions. Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

✉ Lexing Ying
lexing@stanford.edu

[1] Department of Mathematics and ICME, Stanford University, Stanford, CA 94305, USA

Published online: 12 September 2020

The goal of this paper is to develop fast first-order algorithms for identifying minimums of (1). When F is convex (for example, when W is positive-definite), there exists a unique global minimizer and the goal is to compute this global minimizer efficiently. When F is non-convex, there are typically many local minimums and the more moderate goal is to find one such local minimum.

There are several difficulties for computing local minima for (1). First, this is an optimization problem over probability simplex, hence one needs to deal with the constraints $p(x) \geq 0$ and $\int p(x)\mathrm{d}x = 1$. Second, when the reference measure $\mu(x)$ varies drastically for different $x \in \Omega$, the optimization problem can be quite ill-conditioned. Third, we aim to avoid costly second-order Newton or quasi-Newton methods that involve matrix inversions or solves.

1.1 Motivations and Approach

Our approach is motivated by the mirror descent algorithm [16] popularized recently in the machine learning community. Because of several nice computational and analytical features, the mirror descent algorithm has played a significant role in online learning and optimization. For an objective function $E(p)$ over the space of probability densities, it finds a minimizer of $E(p)$ as follows. Given a current density p^k, each step solves for

$$\tilde{p} = \mathrm{argmin}_p\, E(p^k) + \frac{\delta E}{\delta p}(p^k) \cdot (p - p^k) + \frac{1}{\eta} D_{\mathrm{KL}}(p||p^k) \tag{2}$$

and then projects $\tilde{p}$ back to the space of probability densities. Taking derivative of (2) in p results in

$$\eta \frac{\delta E}{\delta p}(p^k) + \ln(\tilde{p}/p^k) + 1 = 0,$$

with $\tilde{p}$ proportional to $p^k \exp\left(-\eta \frac{\delta E}{\delta p}(p^k)\right)$. Projecting it back to the probability simplex via rescaling gives

$$p^{k+1} = \frac{1}{Z} p^k \exp\left(-\eta \frac{\delta E}{\delta p}(p^k)\right), \quad Z = \int p^k \exp\left(-\eta \frac{\delta E}{\delta p}(p^k)\right) \mathrm{d}x. \tag{3}$$

Let us now give a different derivation of the mirror descent algorithm from a more numerical analysis perspective. The starting point is the natural gradient flow of $E(p)$ with the Fisher–Rao metric diag$(1/p)$:

$$\dot{p} = -\frac{1}{1/p}\left(\frac{\delta E}{\delta p} + c\right) = -p\left(\frac{\delta E}{\delta p} + c\right),$$

where $\frac{\delta E}{\delta p}$ is Frechet derivative and c is the Lagrange multiplier associated with $\int_\Omega p(x)\mathrm{d}x = 1$. Moving p to the left hand side gives rise to an equation of $\ln p$.

$$(\dot{\ln p}) = -\left(\frac{\delta E}{\delta p} + c\right).$$

Using the explicit Euler method in the new variable $\ln p$ with step size η results in

$$\ln p^{k+1} = \ln p^k - \eta\left(\frac{\delta E}{\delta p}(p^k) + c\right),$$

where c is determined from the condition $\int p^{k+1} \mathrm{d}x = 1$ and this is equivalent to (3). This derivation shows that mirror descent can be viewed as the explicit Euler discretization of the natural gradient flow in the reparameterization $\phi(p) \equiv \ln p$.

The mirror descent is effective when the Hessian of the energy function $E(p)$ is close to the Fisher–Rao metric $1/p$, up to a constant scaling. This is the case for

$$E(p) = \int p(x) \ln p(x) \mathrm{d}x + \int V(x) p(x) \mathrm{d}x,$$

where the Hessian is exactly the Fisher–Rao metric. In this case, the natural gradient is

$$(\dot{\ln} p) = -(\ln p + V + c).$$

This is a linear system of ordinary differential equations with coefficient 1 in the new variable $\ln p$. The stiffness is gone and one can take large steps.

Coming back to the free energy (1), the mirror descent algorithm described above is not particularly effective, due to the existence of the reference measure μ (in the reverse KL and Hellinger cases) as well as the extra interacting term W. In fact, for general μ and W, the Fisher–Rao metric $1/p$ in the natural gradient algorithm is quite far away from the Hessian matrix of the Newton method. Therefore, there is no reason to expect the standard mirror descent algorithm to be efficient. Our approach consists of the following steps:

- Choose an appropriate diagonal metric based on μ and W;
- Design a reparameterization function ϕ based on the chosen metric;
- Derive the algorithm by performing the explicit Euler discretization;
- Work out the renormalization step.

1.2 Related Work

The mirror descent algorithm [3,16] was proposed as an effective first-order method for convex optimization by taking into consideration the geometry of the problem. For certain types of constraint sets, the mirror descent algorithm is nearly optimal among first order methods [6], offering an almost dimensional independent convergence rate. In the setting of online optimization, mirror descent also allows one to obtain a bound for the cumulative regret [2,5]. There is a vast literature on mirror descent and related algorithms and we refer to [6,20] for further discussions.

The interacting free energy of form (1) appear in several applications, such as Keller–Segel models [18] in mathematical biology, as well as the granular flow in kinetic theory [7,22]. In these applications, the evolution of the probability density is governed by the Wasserstein gradient flow [11,17] of the free energy, i.e., the gradient flow with respect to the Wasserstein metric $-\nabla \cdot (p\nabla(\cdot))$. The main computational task in these applications is to compute the evolution of the Wasserstein gradient flow and several numerical methods based on finite element, finite volume, and particle methods [4,8,12–14] have been proposed for this. Compared with these algorithms, the goal of this paper is different as we only care about the minimizers. Therefore, we have the freedom to pick any descent dynamics that leads to the minimizer. As we have seen, our flow is closer to the natural gradient rather than the Wasserstein gradient.

1.3 Contents

The paper considers three common cases of the divergence term $D(p||\mu)$ and is organized as follows. Section 2 addresses the Kullback–Leibler divergence, Sect. 3 is about the reverse Kullback–Leibler case, and finally Sect. 4 discusses the Hellinger distance case. In each case, we address both the case of positive-definite W term as well as the general situation of non-positive-definite W.

As the metric adopted here is of the Fisher–Rao type as opposed to the Wasserstein type, there is no derivative involved in the computation. To simplify the presentation and also to make connection with the numerical implementation, we work with a probability density $\{p_1, \dots, p_n\}$ over a discrete set of n points $\{x_1, \dots, x_n\}$ rather than over a continuous space. The interacting free energy can be written as

$$F(p) = D(p||\mu) + \sum_i p_i V_i + \frac{1}{2} \sum_{ij} p_i W_{ij} p_j.$$

This is indeed the setup when (1) is discretized with a numerical treatment.

1.4 Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

2 Kullback–Leibler Divergence

For the KL divergence case,

$$D_{\text{KL}}(p||\mu) = \sum_{i=1}^n p_i \ln p_i/\mu_i = \sum_{i=1}^n p_i \ln p_i - \sum_{i=1}^n p_i \ln \mu_i.$$

Since the last term involving μ can be absorbed into the potential V, it is convenient to assume μ to be the uniform measure and consider equivalently

$$F_{\text{KL}}(p) = \sum_i p_i \ln p_i + \sum_i V_i p_i + \frac{1}{2} \sum_{i,j} p_i W_{ij} p_j.$$

The Hessian is given by

$$\frac{\delta^2 F_{\text{KL}}}{\delta p^2} = \text{diag}\left(\frac{1}{p}\right) + W.$$

When W is non-positive-definite, the safe way is to just use $\text{diag}(1/p)$ as the gradient metric. When W is positive-definite, we extract the diagonal $\alpha = \text{diag}(W) \in \mathrm{R}^n$ of W and use $\text{diag}(1/p + \alpha)$ as the gradient metric.

2.1 Non-positive-Definite Case

Using $\text{diag}(1/p)$ as the metric, the gradient flow is

$$\dot{p} = -p(\ln p + V + Wp + c).$$

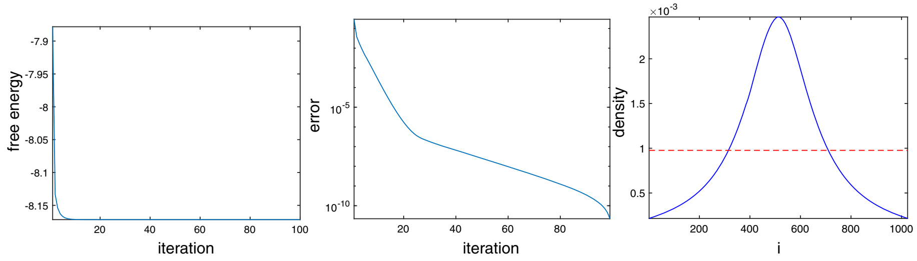

Fig. 1 KL divergence, non-positive-definite case with a Keller–Segel free energy. Left: free energy versus iteration. Middle: free energy error versus iteration. Right: density p at the final iteration (solid curve) compared with the uniform density (dashed curve)

Moving the metric to the left hand side gives

$$(\dot{\ln p}) = -(\ln p + V + Wp + c).$$

If we introduce a reparameterization from $p \in \mathrm{R}^n$ to $g \in \mathrm{R}^n$ with $g_i = \phi_i(p_i) \equiv \ln p_i$ and $p_i = \phi_i^{-1}(g_i) = \exp(g_i)$

$$\begin{aligned} \phi_i &: p_i \to g_i, \quad (0,1) \to (-\infty, 0), \\ \phi_i^{-1} &: g_i \to p_i, \quad (-\infty, 0) \to (0,1), \end{aligned}$$

the gradient flow becomes

$$\dot{g} = -(g + V + Wp + c).$$

The explicit Euler discretization gives

$$\begin{aligned} \tilde{g} &= g^k - \Delta t(g^k + V + Wp^k), \\ g^{k+1} &= \tilde{g} + c. \end{aligned}$$

The constant c is determined by the normalization condition

$$\sum_i \phi_i^{-1}(\tilde{g}_i + c) = 1,$$

which leads to $c = -\ln\left(\sum_i \exp(\tilde{g}_i)\right)$.

We illustrate the efficiency of the algorithm with a Keller–Segel model. Consider the domain $[0,1]$ discretized with $n = 1024$ points $\{x_i = \frac{i}{n}\}$. The potential V is zero and the interacting term is

$$W_{ij} = \frac{3}{2} \ln(|x_i - x_j| + \varepsilon)$$

with $\varepsilon = 10^{-6}$. The step size Δt is taken to be 1. Starting from a random initial condition, we run the descent algorithm for 100 steps. The results are summarized in Fig. 1. At the end of the 100 iterations, the free energy error is of order 10^{-10}. The final density shows the concentration property of the Keller–Segel free energy.

2.2 Positive-Definite Case

Using $\mathrm{diag}(1/p) + \alpha$ as the metric, the gradient flow is

$$\dot{p} = -\frac{1}{1/p + \alpha}(\ln p + V + Wp + c).$$

Moving the metric to the left hand side gives

$$\dot{(\ln p + \alpha p)} = -(\ln p + \alpha p + V + (W - \alpha)p + c).$$

If we introduce a reparameterization from $p \in \mathrm{R}^n$ to $g \in \mathrm{R}^n$ with $g_i = \phi_i(p_i) \equiv \ln(p_i) + \alpha_i p_i$

$$\begin{aligned} \phi_i &: p_i \to g_i, \quad (0, 1) \to (-\infty, \alpha_i), \\ \phi_i^{-1} &: g_i \to p_i, \quad (-\infty, \alpha_i) \to (0, 1), \end{aligned}$$

the gradient flow becomes

$$\dot{g} = -(g + V + (W - \alpha)p + c).$$

The explicit Euler discretization gives

$$\begin{aligned} \tilde{g} &= g^k - \Delta t(g^k + V + (W - \alpha)p^k), \\ g^{k+1} &= \tilde{g} + c. \end{aligned}$$

The constant c is determined by the normalization condition

$$\sum_i \phi_i^{-1}(\tilde{g}_i + c) = 1.$$

Let us observe that $\sum_i \phi_i^{-1}(\tilde{g}_i + c)$ is an increasing function in c as each ϕ_i^{-1} is increasing. The correct value c can be shown to be in

$$\left(\min\left(\ln\frac{1}{n} + \frac{\alpha_i}{n} - \tilde{g}_i\right), \min(\alpha_i - \tilde{g}_i)\right).$$

Plugging the two endpoints of the interval shows that at the left endpoint $\sum_i \phi_i^{-1}(\tilde{g}_i + c) < 1$ and at the right endpoint $\sum_i \phi_i^{-1}(\tilde{g}_i + c) > 1$. Therefore, there is a unique c value satisfies $\sum_i \phi_i^{-1}(\tilde{g}_i + c) = 1$ within this interval. This can be easily found using Newton, bisection, or interpolation methods [10].

To illustrate the efficiency of this algorithm, we consider the periodic domain [0, 1] discretized with $n = 1024$ points. The potential V is chosen to be $V_i = \sin(4\pi x_i)$ and the interacting term is

$$W_{ij} = \begin{cases} \alpha, & i = j, \\ \alpha/2, & i = j \pm 1, \\ 0, & \text{otherwise}, \end{cases}$$

with $\alpha = 10^3$, leading to $\alpha_i = 10^3$ for each i. The step size Δt is taken to be 1. Starting from a random initial condition, we run the algorithm for 100 steps with the results summarized in Fig. 2. Within 20 iterations, it reaches within 10^{-15} accuracy. The final probability density shows that the interacting term in the free energy further suppresses oscillations in the minimizing density.

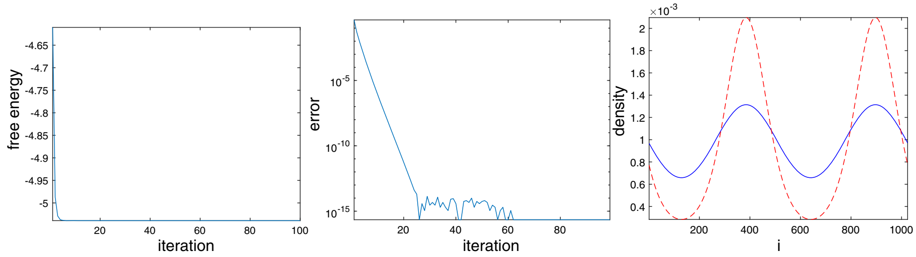

Fig. 2 KL divergence, positive-definite case. Left: free energy versus iteration. Middle: free energy error versus iteration. Right: density p at the final iteration (solid curve) compared with the minimizing density $\exp(-V_i)/Z$ if W is set to zero (dashed curve)

3 Reverse Kullback–Leibler Divergence

For the reverse KL divergence

$$D_{\mathrm{rKL}}(p||\mu) = \sum_i \mu_i \ln \mu_i/p_i = \sum_i \mu_i \ln \mu_i - \sum_i \mu_i \ln p_i.$$

The free energy is now

$$F_{\mathrm{rKL}}(p) = -\sum_i \mu_i \ln p_i + \sum_i V_i p_i + \frac{1}{2}\sum_{i,j} p_i W_{ij} p_j.$$

The Hessian is given by

$$\frac{\delta^2 F_{\mathrm{rKL}}}{\delta p^2} = \operatorname{diag}\left(\frac{\mu}{p^2}\right) + W$$

and it can be quite far from the mirror descent choice $\operatorname{diag}\left(1/p^2\right)$ even when W is zero, since μ can be drastically different for different i. When W is non-positive-definite, it is safe to continue using $\operatorname{diag}\left(\mu/p^2\right)$ as the gradient metric. When W is positive-definite, we extract the diagonal $\alpha = \operatorname{diag}(W)$ of W and use $\operatorname{diag}\left(\mu/p^2 + \alpha\right)$ as the gradient metric.

3.1 Non-positive-Definite Case

Using $\operatorname{diag}\left(\mu/p^2\right)$ as the metric, the gradient flow is

$$\dot{p} = -\frac{1}{\mu/p^2}\left(-\frac{\mu}{p} + V + Wp + c\right).$$

Moving the metric to the left hand side gives

$$(-\dot{\mu/p}) = -(-\mu/p + V + Wp + c).$$

If we introduce a reparameterization from $p \in \mathrm{R}^n$ to $g \in \mathrm{R}^n$ with $g_i = \phi_i(p_i) \equiv -\mu_i/p_i$ and $p_i = \phi_i^{-1}(g_i) = -\mu_i/g_i$

$$\phi_i : p_i \to g_i, \quad (0,1) \to (-\infty, -\mu_i),$$
$$\phi_i^{-1} : g_i \to p_i, \quad (-\infty, -\mu_i) \to (0,1),$$

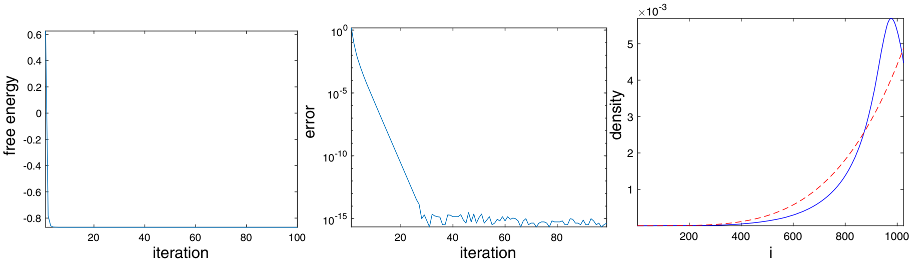

Fig. 3 Reverse KL divergence, non-positive-definite case with a Keller–Segel free energy. Left: free energy versus iteration. Middle: free energy error versus iteration. Right: density p at the final iteration (solid curve) and the reference measure μ (dashed curve)

the gradient flow becomes

$$\dot{g} = -(g + V + Wp + c).$$

The explicit Euler discretization gives

$$\tilde{g} = g^k - \Delta t (g^k + V + Wp^k),$$
$$g^{k+1} = \tilde{g} + c.$$

The constant c is determined by the normalization condition

$$\sum_i \phi_i^{-1}(\tilde{g}_i + c) = 1,$$

Since each ϕ_i^{-1} is increasing, $\sum_i \phi_i^{-1}(\tilde{g}_i + c)$ is an increasing function in c. We claim that the correct value c can be shown to be in

$$(\min(-\tilde{g}_i - n\mu_i), \min(-\tilde{g}_i - \mu_i)).$$

Plugging the two endpoints of the interval shows that at the left endpoint $\sum_i \phi_i^{-1}(\tilde{g}_i + c) < 1$ and at the right endpoint $\sum_i \phi_i^{-1}(\tilde{g}_i + c) > 1$. Therefore, there is a unique c value satisfies $\sum_i \phi_i^{-1}(\tilde{g}_i + c) = 1$ within this interval.

As a numerical example, we consider a Keller–Segel model. Consider the domain [0, 1] discretized with $n = 1024$ points $\{x_i = \frac{i}{n}\}$. The potential V is equal to zero and the interacting term W_{ij} is given by

$$W_{ij} = \frac{2}{3} \ln(|x_i - x_j| + \varepsilon)$$

with $\varepsilon = 10^{-6}$. The reference measure μ is taken to be $\mu_i \sim x_i^4$, leading to a ratio of 10^{12} between the largest and the smallest μ_i values. The step size Δt is taken to be 1. Starting from a random initial condition, we run the descent algorithm for 100 steps and the results are summarized in Fig. 3. Within 30 iterations, the algorithm reaches within 10^{-15} accuracy.

3.2 Positive-Definite Case

Using $\operatorname{diag}\left(\mu/p^2 + \alpha\right)$ as the metric, the gradient flow is

$$\dot{p} = -\frac{1}{\mu/p^2 + \alpha}(\ln p + V + Wp + c).$$

Moving the metric to the left hand side gives

$$\dot{(-\mu/p+\alpha p)} = -(-\mu/p+\alpha p + V + (W-\alpha)p + c).$$

If we introduce a reparameterization from $p \in \mathrm{R}^n$ to $g \in \mathrm{R}^n$ with $g_i = \phi_i(p_i) \equiv -\mu_i/p_i + \alpha_i p_i$ and $p_i = \phi_i^{-1}(g_i) = \frac{g_i+\sqrt{g_i^2+4\alpha_i\mu_i}}{2\alpha_i}$

$$\begin{aligned} &\phi_i : p_i \rightarrow g_i, \quad (0,1) \rightarrow (-\infty, -\mu_i+\alpha_i), \\ &\phi_i^{-1} : g_i \rightarrow p_i, \quad (-\infty, -\mu_i+\alpha_i) \rightarrow (0,1), \end{aligned}$$

the gradient flow becomes

$$\dot{g} = -(g + V + (W-\alpha)p + c).$$

The Explicit Euler discretization gives

$$\begin{aligned} &\tilde{g} = g^k - \Delta t(g^k + V + (W-\alpha)p^k), \\ &g^{k+1} = \tilde{g} + c. \end{aligned}$$

The constant c is determined by the normalization condition

$$\sum_i \phi_i^{-1}(\tilde{g}_i + c) = 1,$$

which can be solved since it is monotone. The correct value c can be shown to be in

$$\left(\min\left(-\tilde{g}_i - n\mu_i + \frac{\alpha_i}{n}\right), \min(-\tilde{g}_i - \mu_i + \alpha_i)\right).$$

Plugging the two endpoints of the interval shows that the left endpoint $\sum_i \phi_i^{-1}(\tilde{g}_i + c) < 1$ and at the right endpoint $\sum_i \phi_i^{-1}(\tilde{g}_i + c) > 1$. Therefore, there is a unique c value satisfies $\sum_i \phi_i^{-1}(\tilde{g}_i + c) = 1$ within this interval.

As a numerical example, consider the periodic domain $[0, 1]$ discretized with $n = 1024$ points. The potential V is chosen to be zero and the interacting term is

$$W_{ij} = \begin{cases} \alpha, & i = j, \\ \alpha/2, & i = j \pm 1, \\ 0, & \text{otherwise}, \end{cases}$$

with $\alpha = 10^2$, leading to $\alpha_i = 10^2$ for each i. The reference measure μ is taken to be $\mu_i \sim x_i^3$. The step size Δt is taken to be 1. Starting from a random initial condition, we run the descent algorithm for 100 steps with the results summarized in Fig. 4. After about only 10 iterations, the free energy error is reduced to about 10^{-15} .

4 Hellinger Divergence

For the Hellinger divergence

$$D_{\mathrm{H}}(p||\mu) = \sum_i (\sqrt{p_i} - \sqrt{\mu_i})^2 = -2\sum_i \sqrt{\mu_i p_i} + \text{cst}.$$

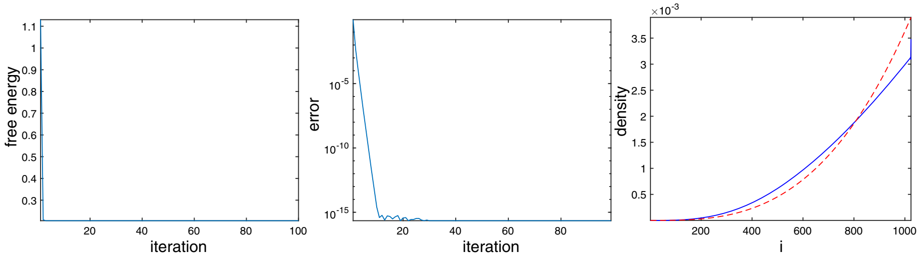

Fig. 4 Reverse KL divergence, positive-definite case. Left: free energy versus iteration. Middle: free energy error versus iteration. Right: density p at the final iteration (solid curve) and the reference measure μ (dashed curve)

The free energy up to a constant is

$$F_{\mathrm{H}}(p) = -2\sum_i \sqrt{\mu_i p_i} + \sum_i V_i p_i + \frac{1}{2}\sum_{i,j} p_i W_{ij} p_j.$$

The Hessian is given by

$$\frac{\delta^2 F_{\mathrm{H}}}{\delta p^2} = \mathrm{diag}\left(\frac{\mu^{1/2}}{2p^{3/2}}\right) + W.$$

Notice that the Hessian can be quite far from the mirror descent choice $\mathrm{diag}(1/(2p^{3/2}))$ even when W is zero, since μ can be drastically different for different i. When W is non-positive-definite, it is safe to continue using $\mathrm{diag}\left(\mu^{1/2}/(2p^{3/2})\right)$ as the gradient metric. When W is positive-definite, we extract the diagonal $\alpha = \mathrm{diag}(W)$ and use $\mathrm{diag}\left(\mu^{1/2}/(2p^{3/2}) + \alpha\right)$ as the gradient metric.

4.1 Non-positive-Definite Case

Using $\mathrm{diag}\left(\mu^{1/2}/(2p^{3/2})\right)$ as the metric, the gradient flow is

$$\dot{p} = -\frac{1}{\mu^{1/2}/(2p^{3/2})}\left(-\sqrt{\frac{\mu}{p}} + V + Wp + c\right).$$

Moving the metric to the left hand side gives

$$\dot{\left(-\sqrt{\mu/p}\right)} = -(-\sqrt{\mu/p} + V + Wp + c).$$

If we introduce a reparameterization from $p \in \mathrm{R}^n$ to $g \in \mathrm{R}^n$ with $g_i = \phi_i(p_i) \equiv -\sqrt{\mu_i/p_i}$ and $p_i = \phi_i^{-1}(g_i) = \mu_i/g_i^2$

$$\begin{aligned} \phi_i &: p_i \to g_i, \quad (0,1) \to (-\infty, -\sqrt{\mu_i}), \\ \phi_i^{-1} &: g_i \to p_i, \quad (-\infty, -\sqrt{\mu_i}) \to (0,1), \end{aligned}$$

the gradient flow becomes

$$\dot{g} = -(g + V + Wp + c).$$

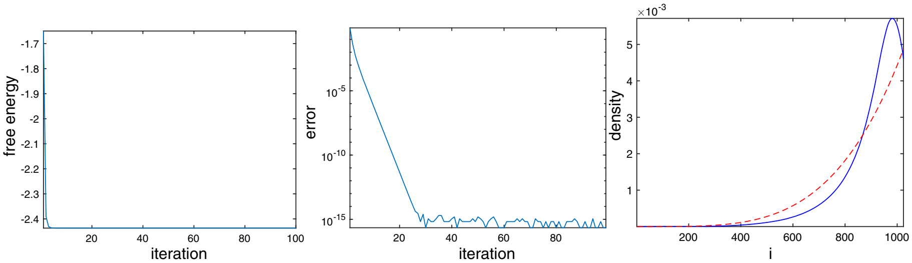

Fig. 5 Hellinger divergence, non-positive-definite case with a Keller–Segel free energy. Left: free energy versus iteration. Middle: free energy error versus iteration. Right: density p at the final iteration (solid curve) and the reference measure μ (dashed line)

The explicit Euler discretization gives

$$\begin{aligned}\tilde{g} &= g^k - \Delta t(g^k + V + Wp^k),\\ g^{k+1} &= \tilde{g} + c.\end{aligned}$$

The constant c is determined by the normalization condition

$$\sum_i \phi_i^{-1}(\tilde{g}_i + c) = 1,$$

which can be solved since it is monotone. The correct value c can be shown to be in

$$\left(\min\left(-\tilde{g}_i - \sqrt{n\mu_i}\right), \min(-\tilde{g}_i - \sqrt{\mu_i})\right).$$

Plugging the two endpoints of the interval shows that at the left endpoint $\sum_i \phi_i^{-1}(\tilde{g}_i + c) < 1$ and at the right endpoint $\sum_i \phi_i^{-1}(\tilde{g}_i + c) > 1$. Therefore, there is a unique c value satisfies $\sum_i \phi_i^{-1}(\tilde{g}_i + c) = 1$ within this interval.

We illustrate the efficiency of the algorithm using a Keller–Segel model. Consider the domain [0, 1] discretized with $n = 1024$ points $\{x_i = \frac{i}{n}\}$. The potential V is zero and the interacting term W_{ij} is given by

$$W_{ij} = \frac{1}{3}\ln(|x_i - x_j| + \varepsilon)$$

with $\varepsilon = 10^{-6}$. The reference measure μ is taken to be $\mu_i \sim x_i^4$. The step size Δt is taken to be 1. Starting from a random initial condition, we run the descent algorithm for 100 steps and the results are summarized in Fig. 5. Within 30 iterations, it reaches within 10^{-15} accuracy.

4.2 Positive-Definite Case

Using diag $\left(\mu^{1/2}/(2p^{3/2}) + \alpha\right)$ as the metric, the gradient flow is

$$\dot{p} = -\frac{1}{\mu^{1/2}/(2p^{3/2}) + \alpha}\left(-\sqrt{\frac{\mu}{p}} + V + Wp + c\right).$$

Moving the metric to the left hand side gives

$$\dot{\left(-\sqrt{\mu/p} + \alpha p\right)} = -(-\sqrt{\mu/p} + \alpha p + V + (W - \alpha)p + c).$$

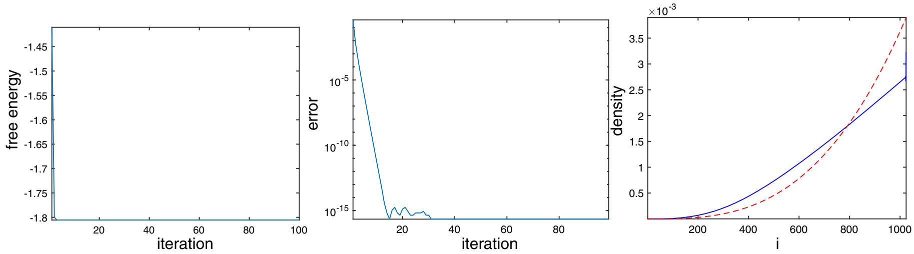

Fig. 6 Hellinger divergence, positive-definite case. Left: free energy versus iteration. Middle: free energy error versus iteration. Right: density p at the final iteration (solid curve) and the reference measure μ (dashed curve)

If we introduce a reparameterization from $p \in \mathrm{R}^n$ to $g \in \mathrm{R}^n$ with $g_i = \phi_i(p_i) \equiv -\sqrt{\mu_i/p_i} + \alpha_i p_i$:

$$\begin{aligned} \phi_i : p_i \to g_i, \quad & (0,1) \to (-\infty, -\sqrt{\mu_i} + \alpha_i), \\ \phi_i^{-1} : g_i \to p_i, \quad & (-\infty, -\sqrt{\mu_i} + \alpha_i) \to (0,1), \end{aligned}$$

the gradient flow becomes

$$\dot{g} = -(g + V + (W - \alpha)p + c).$$

An explicit Euler discretization gives

$$\begin{aligned} \tilde{g} &= g^k - \Delta t (g^k + V + (W - \alpha)p^k), \\ g^{k+1} &= \tilde{g} + c. \end{aligned}$$

The constant c is determined by the normalization condition

$$\sum_i \phi_i^{-1}(\tilde{g}_i + c) = 1,$$

which can be solved since it is monotone. The correct value c can be shown to be in

$$\left(\min\left(-\tilde{g}_i - \sqrt{n\mu_i} + \frac{\alpha_i}{n} \right), \min(-\tilde{g}_i - \sqrt{\mu_i} + \alpha_i) \right).$$

Plugging the two endpoints of the interval shows that the left endpoint $\sum_i \phi_i^{-1}(\tilde{g}_i + c) < 1$ and at the right endpoint $\sum_i \phi_i^{-1}(\tilde{g}_i + c) > 1$. Therefore, there is a unique c value satisfies $\sum_i \phi_i^{-1}(\tilde{g}_i + c) = 1$ within this interval.

In the numerical test, we consider the periodic domain $[0, 1]$ discretized with $n = 1024$ points. The potential V is chosen to be zero and the interacting term is

$$W_{ij} = \begin{cases} \alpha, & i = j, \\ \alpha/2, & i = j \pm 1, \\ 0, & \text{otherwise}, \end{cases}$$

with $\alpha = 10^2$. This leads to $\alpha_i = 10^2$ for each $i = 1, \ldots, n$. The reference measure μ is chosen such that $\mu_i \sim x_i^3$. The step size Δt is taken to be 1. Starting from a random initial condition, we run the descent algorithm for 100 steps. The results are summarized in Fig. 6. Within about 15 iterations, it converges to an accuracy of order 10^{-15}.

5 Discussions

This paper proposes mirror-descent-type algorithms for minimizing interacting free energies. Below we point out a few questions for future work. First, the proposed algorithms are obtained from discretizing the continuous-time gradient flow with a new metric based on μ and W. One can also derive the algorithm in a more traditional mirror descent form by starting from the corresponding Bregman divergences.

Second, this paper considers three cases: KL divergence, reverse KL divergence, and Hellinger divergence. In fact, the same procedure can be extended to most α-divergences [1].

When we treat the non-positive-definite case, W is simply dropped in the design of the new metric. A more accurate, but potentially more computationally intensive, alternative is to find a positive-definite approximation to W and then combine it with the Hessian from the divergence term.

This interacting term of the free energy considered in this paper is only of quadratic form. It is plausible that a similar procedure can be developed for non-quadratic interacting terms, as long as there is an efficient way to approximate the diagonal of the Hessian.

References

1. Amari, S.: Information Geometry and Its Applications, vol. 194. Springer, Berlin (2016)
2. Arora, S., Hazan, E., Kale, S.: The multiplicative weights update method: a meta-algorithm and applications. Theory Comput. **8**(1), 121164 (2012)
3. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. **31**(3), 167175 (2003)
4. Bessemoulin-Chatard, M., Filbet, F.: A finite volume scheme for nonlinear degenerate parabolic equations. SIAM J. Sci. Comput. **34**(5), B559–B583 (2012)
5. Bubeck, S.: Introduction to online optimization. Lect. Notes **2** (2011)
6. Bubeck, S., et al.: Convex optimization: algorithms and complexity. Found. Trends R Mach. Learn. **8**(3–4), 231357 (2015)
7. Carrillo, J.A., McCann, R.J., Villani, C., et al.: Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. Rev. Mat. Iberoam. **19**(3), 9711018 (2003)
8. Carrillo, J.A., Craig, K., Patacchini, F.S.: A blob method for diffusion. Cal. Var. Partial Differ. Equ. **58**(2), 53 (2019)
9. Chizat, L., Bach, F.: On the global convergence of gradient descent for over-parameterized models using optimal transport. In: Advances in Neural Information Processing Systems, pp. 3036–3046 (2018)
10. Forsythe, G.E., Malcolm, M.A., Moler, C.B.: Computer Methods for Mathematical Computations, vol. 259. Prentice-Hall, Englewood Cliffs (1977)
11. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the fokker-planck equation. SIAM J. Math. Anal. **29**(1), 117 (1998)
12. Li, W., Lu, J., Wang, L.: Fisher information regularization schemes for wasserstein gradientows (2019). arXiv:1907.02152
13. Li, W., Montúfar, G.: Natural gradient via optimal transport. Inf. Geom. **1**(2), 181214 (2018)
14. Liu, J.-G., Wang, L., Zhou, Z.: Positivity-preserving and asymptotic preserving method for 2d Keller–Segal equations. Math. Comput. **87**(311), 11651189 (2018)
15. Mei, S., Montanari, A., Nguyen, P.-M.: A mean field view of the landscape of two-layer neural networks. Proc. Natl. Acad. Sci. **115**(33), E7665–E7671 (2018)
16. Nemirovsky, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization. A Wiley-Interscience Publication. Wiley, New York (1983). Translated from the Russian and with a preface by Dawson, E.R. Wiley-Interscience Series in Discrete Mathematics. MR702836
17. Otto, F.: The Geometry of Dissipative Evolution Equations: The Porous Medium Equation. Taylor & Francis, London (2001)
18. Perthame, B.: Transport Equations in Biology. Springer, Berlin (2006)

19. Rotskoff, G.M., Vanden-Eijnden, E.: Neural networks as interacting particle systems: asymptotic convexity of the loss landscape and universal scaling of the approximation error (2018). arXiv:1805.00915
20. Shalev-Shwartz, S., et al.: Online learning and online convex optimization. Found. Trends R Mach. Learn. **4**(2), 107194 (2012)
21. Sirignano, J., Spiliopoulos, K.: Mean field analysis of neural networks (2018). arXiv:1805.01053
22. Villani, C.: Mathematics of granular materials. J. Stat. Phys. **124**(2–4), 781822 (2006)