Some natural gradient algorithms for optimizing functionals over probabilities

Lexing Ying
Department of Mathematics
Stanford University

Oct 29, 2021
Dynamics and Discretization: PDEs, Sampling, and Optimization
Simons Institute, UC Berkeley

Introduction

Consider the task

$$\min_p E(p)$$

- $p(x)$: a probability distribution over $x \in \Omega$
- $E(p)$: a functional of probability distribution

Focus on two problems

- 1. Interacting free energy

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

- 2. Mixed form

$$E(p) = \text{Wasserstein-dist}^2 + \text{KL} + \text{Mahalanobis-dist}^2$$

Discuss new algorithms based on natural gradient

Problem 1

Minimizing the *interacting* free energy over probabilities over Ω

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

where $D(p||\mu)$ divergence, μ ref. measure, W symmetric (the interacting term)

Applications

- Keller-Segel models in biology and granular flows in kinetic theory
- Mean field modeling of neural network training

Goal

- Find a minimum (global/local depending on the convexity of W)
- First order method

Mirror descent

The usual derivation

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

where $D_{\mathrm{KL}}(\cdot||\cdot)$ is the KL divergence.

Taking derivative wrt p

$$\eta \frac{\delta E}{\delta p}(p^k) + \ln\left(p^{k+1}/p^k\right) = \text{cst} \quad \Rightarrow \quad p^{k+1} \propto p^k \exp\left(-\eta \frac{\delta E}{\delta p}(p^k)\right)$$

Applications in statistics, online learning, etc.

Efficiency

MD with D_{KL} is effective if $\frac{\delta^2 E}{\delta p^2} \sim \mathrm{diag}(1/p)$, e.g.

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

MD with D_{KL} is not effective if $\frac{\delta^2 E}{\delta p^2}$ is far from $\mathrm{diag}(1/p)$, e.g.

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

due to

- general D and μ
- the interacting term W

Need new algorithms

Alternative derivation of MD

- 1. Natural gradient with Fisher-Rao metric $\mathrm{diag}(1/p)$:

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

- 2. Moving p to the LHS gives an equation of $\phi(p) = \ln p$.

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

- 3. Explicit Euler discretization with $\Delta t = \eta$

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

- 4. Renormalization

$$p^{k+1} = \frac{1}{Z} p^k \exp\left(-\eta \frac{\delta E}{\delta p}(p^k)\right)$$

This works well for $E(p) = \int p(x) \ln p(x) \mathrm{d}x + \int V(x) p(x) \mathrm{d}x$ because

- $\frac{\delta^2 E}{\delta p^2} \sim \mathrm{diag}(1/p)$.
- This is the Newton flow!

Plan for the general case

- 1. Choose a diagonal metric based on D, μ, and W
- 2. Introduce new $\phi(p)$ and rewrite the flow in $\phi(p)$
- 3. Discretize the $\phi(p)$ equation with explicit Euler
- 4. Work out the renormalization

In the language of MD

- The regularizer should depend on D, μ, and W

Discrete setting: $p = (p_1, \ldots, p_n)$ over point set $\{x_1, \ldots, x_n\}$

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

Diagonal metric

- When W is SPD, use $\frac{\delta^2 D}{\delta p^2} + \mathrm{diag}(w)$ where $w = \mathrm{diag}(W) \in \mathbb{R}^n$
- When W is not SPD, simply use $\frac{\delta^2 D}{\delta p^2}$

In what follows, assume W is SPD

(A) Kullback-Leibler divergence

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

The interacting free energy is

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

The Hessian is given by

$$\frac{\delta^2 E_{\mathrm{KL}}}{\delta p^2} = \mathrm{diag}\left(\frac{1}{p}\right) + W \approx \mathrm{diag}\left(\frac{1}{p} + w\right)$$

1. Using $\mathrm{diag}(1/p + w)$ as the metric

$$\dot{p} = -\frac{1}{1/p + w}(\ln p + V + Wp + c) \Rightarrow (\ln p \dot{+} wp) = -(\ln p + wp + V + (W - w)p + c)$$

2. Introduce variable $g \in \mathbb{R}^n$ with $g_i = \phi_i(p_i) \equiv \ln(p_i) + w_i p_i$

$$\dot{g} = -(g + V + (W - w)p + c), \quad p = \phi^{-1}(g)$$

3. Explicit Euler gives

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

4. c is determined by the normalization condition

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

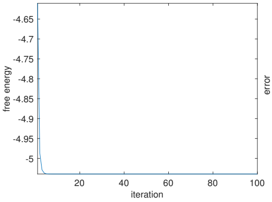
since ϕ_i^{-1} is monotone. In fact $c \in \left(\min_i \left(\ln \frac{1}{n} + \frac{w_i}{n} - \tilde{g}_i\right), \min_i(w_i - \tilde{g}_i)\right)$.

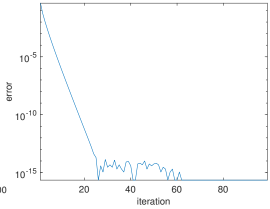
Example

Consider the periodic domain $[0,1]$ discretized with $n = 1024$ points. $V_i = \sin(4\pi x_i)$.

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

with $\alpha = 10^3$. $\Delta t = 1$.





(B) Reverse Kullback-Leibler 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 interacting free energy is

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

The Hessian is given by

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

1. Using $\mathrm{diag}\left(\mu/p^2 + w\right)$ as the metric

$$\dot{p} = \frac{-1}{\mu/p^2 + w}(\ln p + V + Wp + c) \Rightarrow (-\mu/p \dot{+} wp) = -(-\mu/p + wp + V + (W - w)p + c)$$

2. Introduce variable $g \in \mathbb{R}^n$ with $g_i = \phi_i(p_i) \equiv -\mu_i/p_i + w_i p_i$ and $p_i = \phi_i^{-1}(g_i) = \frac{g_i + \sqrt{g_i^2 + 4w_i\mu_i}}{2w_i}$

$$\dot{g} = -(g + V + (W - w)p + c), \quad p = \phi^{-1}(g)$$

3. Explicit Euler gives

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

4. c is determined by the normalization condition

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

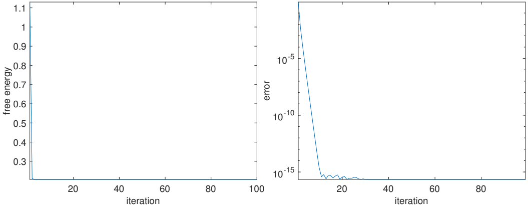
since it is monotone. $c \in \left(\min_i \left(-\tilde{g}_i - n\mu_i + \frac{w_i}{n}\right), \min_i(-\tilde{g}_i - \mu_i + w_i)\right)$.

Example

Consider the periodic domain $[0,1]$ discretized with $n = 1024$ points. $V_i = 0$.

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

with $\alpha = 10^2$. The reference measure $\mu_i \sim x_i^3$. $\Delta t = 1$.



(C) Hellinger divergence

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

The interacting free energy is

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

The Hessian is given by

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

1. Using $\mathrm{diag}\left(\mu^{1/2}/(2p^{3/2})+w\right)$ as the metric

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

2. Introduce variable $g \in \mathbb{R}^n$ with $g_i = \phi_i(p_i) \equiv -\sqrt{\mu_i/p_i}+w_ip_i$:

$$\dot{g} = -(g+V+(W-w)p+c), \quad p = \phi^{-1}(g)$$

3. Explicit Euler gives

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

4. c is determined by the normalization condition

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

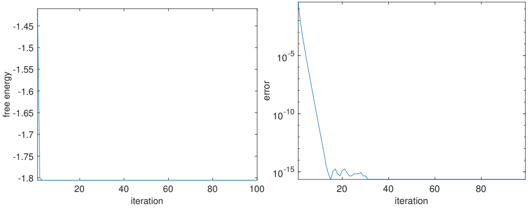
since it is monotone. $c \in \left(\min_i\left(-\tilde{g}_i-\sqrt{n\mu_i}+\frac{w_i}{n}\right), \min_i(-\tilde{g}_i-\sqrt{\mu_i}+w_i)\right).$

Example

Consider the periodic domain $[0, 1]$ discretized with $n = 1024$ points. $V_i = 0$.

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

with $\alpha = 10^2$. The reference measure $\mu \sim x_i^3$. $\Delta t = 1$.



Summary of Problem 1

Natural gradient algorithm for interacting free energies

- 1. Choose a diagonal metric based on D, μ, and W
- 2. Introduce new variable $\phi(p)$ and rewrite the flow in $\phi(p)$
- 3. Discretize the $\phi(p)$ equation with explicit Euler
- 4. Work out the renormalization

Key points

- Newton flow + diagonal Hessian approximation
- The numerical analysis perspective of MD can be useful
- In the language of MD, the regularizer should depend on D, μ, and W

Questions

- High dimensional case

Problem 2

Minimizing $E(p)$ over the probability densities p over Ω

$$E(p) = \alpha_1 E_1(p) + \alpha_2 E_2(p) + \alpha_3 E_3(p),$$

where $\alpha_1, \alpha_2, \alpha_3 \geq 0$ and

- $E_1(p) \sim$ Wasserstein distance square from a base point
- $E_2(p) \sim$ KL divergence $D_{\mathrm{KL}}(p||q) = \int p(x) \ln \frac{p(x)}{q(x)} \mathrm{d}x$.
- $E_3(p) \sim$ Mahalanobis distance square $\frac{1}{2}(p - \mu, A(p - \mu))$ with pseudodifferential A, e.g. $A = (-\Delta)^\beta$

Related to

- Optimal transport
- Maximum mean discrepancy

Goals:

- First order method for minimization
- Low cost

Natural gradient

Choose a metric $\approx$ Hessian of $E(p)$

- Wasserstein GD for $E_1(p) \sim$ Wasserstein distance square

$$\frac{\delta^2 E_1}{\delta p^2}(p) \approx (-\nabla \cdot (p\nabla))^+, \quad \text{metric}^{-1} = -\nabla \cdot (p\nabla)$$

- Fisher-Rao (KL) GD for $E_2(p) \sim$ KL divergence

$$\frac{\delta^2 E_2}{\delta p^2}(p) \approx \mathrm{diag}\left(\frac{1}{p}\right), \quad \text{metric}^{-1} = \mathrm{diag}(p)$$

- Mahalanobis GD for $E_3(p) \sim$ Mahalanobis distance square

$$\frac{\delta^2 E_3}{\delta p^2}(p) \approx A, \quad \text{metric}^{-1} = A^{-1}$$

$\frac{\delta^2 E}{\delta p^2}(p) = \alpha_1 \frac{\delta^2 E_1}{\delta p^2}(p) + \alpha_2 \frac{\delta^2 E_2}{\delta p^2}(p) + \alpha_3 \frac{\delta^2 E_3}{\delta p^2}(p)$ and $\text{metric}^{-1} \approx \left(\frac{\delta^2 E}{\delta p^2}(p)\right)^{-1}$

- None of the three metrics is close to it
- Find a basis that diagonalizes $\frac{\delta^2 E_1}{\delta p^2}(p)$, $\frac{\delta^2 E_2}{\delta p^2}(p)$, $\frac{\delta^2 E_3}{\delta p^2}(p)$
- $\Rightarrow$ Wavelet?

Algorithm

Consider $\Omega = [0,1]^d$ with periodic BC

$$E(p) = \alpha_1 E_1(p) + \alpha_2 E_2(p) + \alpha_3 E_3(p)$$

$$\frac{\delta^2 E}{\delta p^2}(p) = \alpha_1(-\nabla \cdot (p\nabla))^+ + \alpha_2 \operatorname{diag}\left(\frac{1}{p}\right) + \alpha_3 A.$$

Discretized with a uniform grid and denote

- D as discrete differentiation
- W as discrete wavelet transform

$$W^\mathsf{T} \frac{\delta^2 E}{\delta p^2}(p) W \approx \alpha_1 W^\mathsf{T} \left(D^\mathsf{T} \operatorname{diag}(p) D\right)^+ W + \alpha_2 W^\mathsf{T} \operatorname{diag}\left(\frac{1}{p}\right) W + \alpha_3 W^\mathsf{T} A W.$$

$$W^\top \frac{\delta^2 E}{\delta p^2}(p) W \approx \alpha_1 W^\top \left(D^\top \operatorname{diag}(p) D\right)^+ W + \alpha_2 W^\top \operatorname{diag}\left(\frac{1}{p}\right) W + \alpha_3 W^\top A W.$$

All three terms approximately diagonalized

1. $(W^\top D^\top \operatorname{diag}(p) D W)_{ii} = \sum_{s \in S} (DW)_{si}^2 p_s \equiv (H_1 p)_i$ for a matrix H_1

$$W^\top \left(D^\top \operatorname{diag}(p) D\right)^+ W \approx \operatorname{diag}\left(\frac{1}{H_1 p}\right).$$

2. $(W^\top \operatorname{diag}(p) W)_{ii} = \sum_{s \in S} W_{si}^2 p_s \equiv (H_2 p)_i$ for a matrix H_2

$$W^\top \operatorname{diag}\left(\frac{1}{p}\right) W \approx \operatorname{diag}\left(\frac{1}{H_2 p}\right).$$

3. Let $h_3 = \operatorname{diag}(W^\top A W)$ (precomputable)

$$W^\top A W \approx \operatorname{diag}(h_3).$$

Putting together

$$W^\top \frac{\delta^2 E}{\delta p^2}(p) W \approx \operatorname{diag}\left(\frac{\alpha_1}{H_1 p} + \frac{\alpha_2}{H_2 p} + \alpha_3 h_3\right),$$

$$\frac{\delta^2 E}{\delta p^2}(p) \approx W \operatorname{diag}\left(\frac{\alpha_1}{H_1 p} + \frac{\alpha_2}{H_2 p} + \alpha_3 h_3\right) W^\top.$$

From $\frac{\delta^2 E}{\delta p^2}(p) \approx W \operatorname{diag}\left(\frac{\alpha_1}{H_1 p} + \frac{\alpha_2}{H_2 p} + \alpha_3 h_3\right) W^\mathsf{T}$

$$\text{metric}^{-1} = \left(\frac{\delta^2 E}{\delta p^2}(p)\right)^{-1} \approx W \operatorname{diag}\left(\frac{1}{\frac{\alpha_1}{H_1 p} + \frac{\alpha_2}{H_2 p} + \alpha_3 h_3}\right) W^\mathsf{T}.$$

$$\text{Natural grad:} \quad \dot{p} = -\left[W \operatorname{diag}\left(\frac{1}{\frac{\alpha_1}{H_1 p} + \frac{\alpha_2}{H_2 p} + \alpha_3 h_3}\right) W^\mathsf{T}\right] \frac{\delta E}{\delta p}(p).$$

Claim

The computational cost of forming and storing the matrices H_1 and H_2 is $O(n \log n)$.

Claim

For a density $p \in \mathbb{R}^n$ with $p_i > 0$, the computational cost of applying the metric $W \operatorname{diag} \frac{1}{\left(\frac{\alpha_1}{H_1 p} + \frac{\alpha_2}{H_2 p} + \alpha_3 h_3\right)} W^\mathsf{T}$ takes $O(n \log n)$ steps.

Time discretization

$$\text{Natural grad:} \quad \dot{p} = -W \operatorname{diag}\left(\frac{1}{\frac{\alpha_1}{H_1 p} + \frac{\alpha_2}{H_2 p} + \alpha_3 h_3}\right) W^\mathsf{T} \frac{\delta E}{\delta p}(p).$$

We use a backtracking line search algorithm with Armijo condition to enforce positivity.

At time step k with p^k (current approximation)

- Introduce

$$s^k = W \operatorname{diag}\left(\frac{1}{\frac{\alpha_1}{H_1 p^k} + \frac{\alpha_2}{H_2 p^k} + \alpha_3 h_3}\right) W^\mathsf{T} \frac{\delta E}{\delta p}(p^k)$$

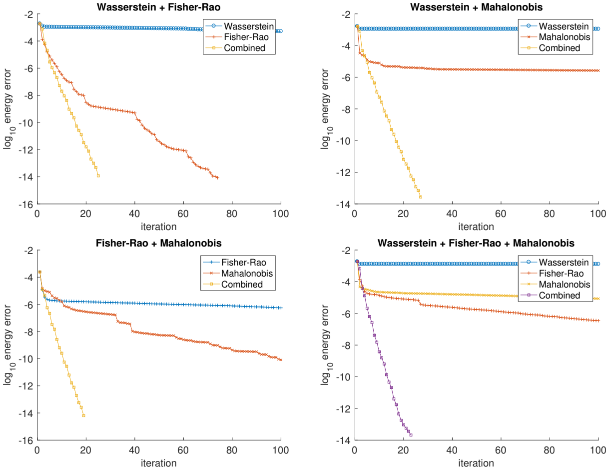
- Starting from $\eta = 1$, halves η repetitively until

$$E(p^k - \eta s^k) - E(p_k) \leq -\frac{1}{2} \eta s^k \cdot \frac{\delta E}{\delta p}(p^k)$$

- Set $p^{k+1} = p^k - \eta s^k$

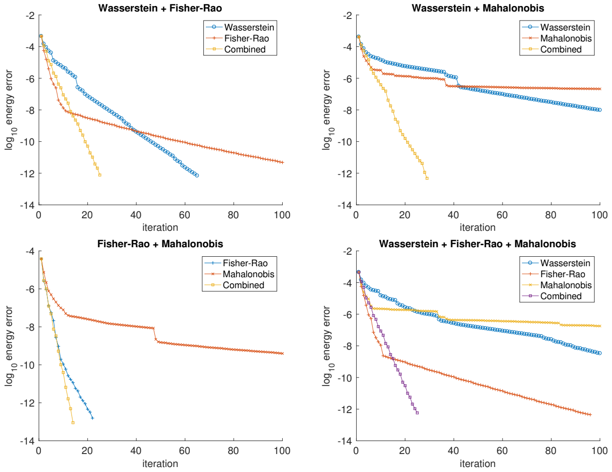
1D example

$$E_1(p) = \frac{1}{2}\|p-\mu\|^2_{\dot{H}^{-1}(\mu)}, \;\; E_2(p) = \sum_s p_s \log\frac{p_s}{\mu_s}, \;\; E_3(p) = \frac{1}{2}(p-\mu)^\mathsf{T}(-\Delta)(p-\mu)$$



2D example

$$E_1(p) = \frac{1}{2}\|p-\mu\|^2_{\dot{H}^{-1}(\mu)},\ \ E_2(p) = \sum_s p_s \log\frac{p_s}{\mu_s},\ \ E_3(p) = \frac{1}{2}(p-\mu)^{\mathsf{T}}(-\Delta)(p-\mu)$$



Summary of Problem 2

Natural gradient for Wasserstein + Fisher-Rao + Mahalanobis

- Diagonal Hessian approximation in wavelet basis
- Backtracking line search

Key-points

- Newton flow + diagonal Hessian approximation
- Harmonic analysis, wavelets

Questions

- Wavelets for general domain
- $p(x) \approx 0$ or non-smooth $p(x)$
- High dimensional case

Thank you

Research supported by NSF and DOE

References

- Mirror Descent Algorithms for Minimizing Interacting Free Energy. Journal of Scientific Computing 84 (2020).
- *Li Wang and Ming Yan*, Hessian informed mirror descent, arXiv:2106.13477
- Natural Gradient for Combined Loss Using Wavelets. Journal of Scientific Computing 86 (2021)