1867

Accelerated Flow For Probability Distributions

Amirhossein Taghvaei, Prashant Mehta

Coordinated Science Laboratory, University of Illinois at Urbana-Champaign

36th International Conference on Machine Lerning (ICML 2019), Long Beach, CA, USA

Motivation and objective

- Many machine learning problems are modelled as an optimization problem on the space of probability distributions
  - Bayesian inference
  - Learning generative models
  - Policy optimization in reinforcement learning
- Solution approaches by constructing gradient flows for probability distributions
  - Liu & Wang, 2016. *"Stein variational gradient descent"*
  - Zhang, et. al. 2018. *"Policy optimization as wasserstein gradient flows"*
  - Frogner & Poggio, 2018. *"Approximate inference with wasserstein gradient flows"*
  - Chizat & Bach, 2018. *"On the global convergence of gradient descent for over-parameterized models using optimal transport"*
- **This paper:** Construct accelerated gradient flows for probability distribution

Approach and main idea

	Euclidean space	Space of probability distributions
	Gradient descent	Wasserstein gradient flow
Accelerated methods		?

- (Wibisono, et. al. 2017) proposed a variational formulation to construct accelerated flows on Euclidean space
- Our approach is to extend the variational formulation for probability distributions

Variational formulation in Euclidean space

Optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) \quad \text{(Assume } f \text{ is convex)}$$

Gradient flow:

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -\nabla f(x_t)$$

Accelerated flow: (Su, et. al. 2014)

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = e^{\alpha_t-\gamma_t}y_t$$
$$\frac{\mathrm{d}y_t}{\mathrm{d}t} = -e^{\alpha_t+\beta_t+\gamma_t}\nabla f(x_t)$$

Variational formulation:

$$\text{Minimize} \quad \int_0^\infty e^{\alpha_t+\gamma_t}\Big(\frac{1}{2}|e^{-\alpha_t}u_t|^2 - e^{\beta_t}f(x_t)\Big)\,\mathrm{d}t$$
$$\text{Subject to} \quad \frac{\mathrm{d}x_t}{\mathrm{d}t} = u_t$$

- Accelerated flow is the solution to the variational problem (Wibisono, et. al. 2017)

Wasserstein gradient flow

Optimization problem:

$$\min_{\rho\in\mathcal{P}_2(\mathbb{R}^d)} \mathsf{F}(\rho) = D(\rho|\rho_\infty) \quad \text{(relative entropy)}$$

Gradient flow: (Jordan, et. al. 1998)

pde form: $\frac{\partial \rho_t}{\partial t} = -\nabla\cdot(\rho_t\log(\rho_\infty)) + \Delta\rho_t,$ (Fokker-Planck eq.)

probabilistic form: $\mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t,$ (Langevin eq.)

where $f = -\log(\rho_\infty)$

Variational formulation for probability distributions

pde form:

$$\text{Minimize} \quad \int_0^\infty e^{\alpha_t+\gamma_t}\Big(\int_{\mathbb{R}^d}\frac{1}{2}|e^{-\alpha_t}u_t(x)|^2\rho_t(x)\,\mathrm{d}x - e^{\beta_t}D(\rho_t|\rho_\infty)\Big)\,\mathrm{d}t$$
$$\text{Subject to} \quad \frac{\partial\rho_t}{\partial t} + \nabla\cdot(\rho_t u_t) = 0$$

probabilistic form:

$$\text{Minimize} \quad \mathsf{E}\Big[\int_0^\infty e^{\alpha_t+\gamma_t}\Big(\frac{1}{2}|e^{-\alpha_t}U_t|^2 - e^{\beta_t}\log(\frac{\rho_t(X_t)}{\rho_\infty(X_t)})\Big)\,\mathrm{d}t\Big]$$
$$\text{Subject to} \quad \frac{\mathrm{d}X_t}{\mathrm{d}t} = U_t$$

- It is a mean-field optimal control problem (Bensoussan, et al. 2013, Carmona & Delarue, 2017)

Main result

Accelerated flow:

pde form:
$$\frac{\partial\rho_t}{\partial t} = -\nabla\cdot(\rho_t e^{\alpha_t-\gamma_t}\nabla\phi_t)$$
$$\frac{\partial\phi_t}{\partial t} = -e^{\alpha_t-\gamma_t}\frac{|\nabla\phi_t|^2}{2} - e^{\alpha_t+\beta_t+\gamma_t}\log(\frac{\rho_t}{\rho_\infty})$$

probabilistic form:
$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = e^{\alpha_t-\gamma_t}Y_t$$
$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = -e^{\alpha_t+\beta_t+\gamma_t}\nabla\log(\frac{\rho_t(X_t)}{\rho_\infty(X_t)})$$

Relationship:

$$\mathrm{Law}(X_t) = \rho_t, \qquad U_t = u_t(X_t), \qquad Y_t = \nabla\phi_t(X_t)$$

Convergence:

- Assume ρ_∞ is log-concave and $d=1$
- Lyapunov function $V(t) = \frac{1}{2}\mathsf{E}[|X_t + e^{-\gamma_t}Y_t - T_{\rho_t}^{\rho_\infty}(X_t)|^2] + e^{\beta_t}D(\rho_t|\rho_\infty)$
- Time derivative $\frac{\mathrm{d}V}{\mathrm{d}t}(t) \le 0$ and

$$D(\rho_t|\rho_\infty) \le O(e^{-\beta_t})$$

Numerical algorithm: Interacting particle system

Simulate N particles $\{(X_t^1,Y_t^1),\ldots,(X_t^N,Y_t^N)\}$

$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = e^{\alpha_t-\gamma_t}Y_t^i, \qquad X_0^i \sim \rho_0$$
$$\frac{\mathrm{d}Y_t^i}{\mathrm{d}t} = -e^{\alpha_t+\beta_t+\gamma_t}(\nabla f(X_t^i) + \underbrace{\nabla\log(\rho_t(X_t^i))}_{\text{interaction term}})$$

Interaction term is approximated with particles

- (parametric) Gaussian approximation

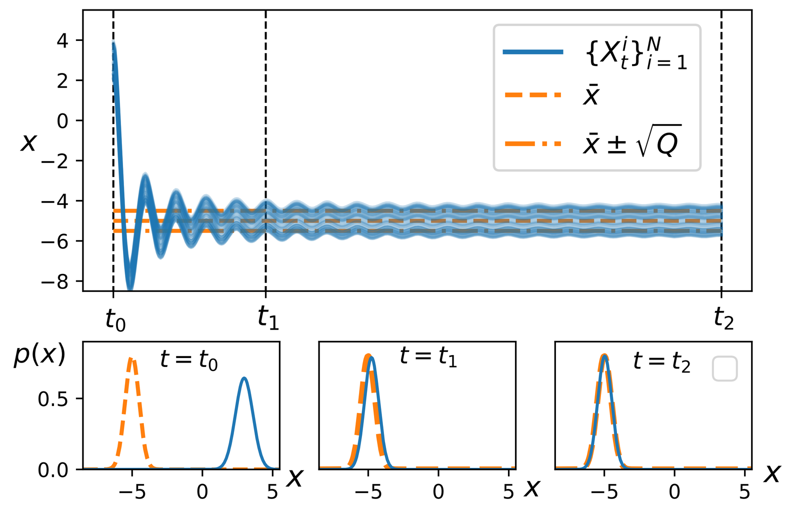
$$\nabla\log(\rho(x)) \approx -\Sigma^{-1}(x-m), \qquad m,\Sigma = \text{empirical mean and covariance}$$

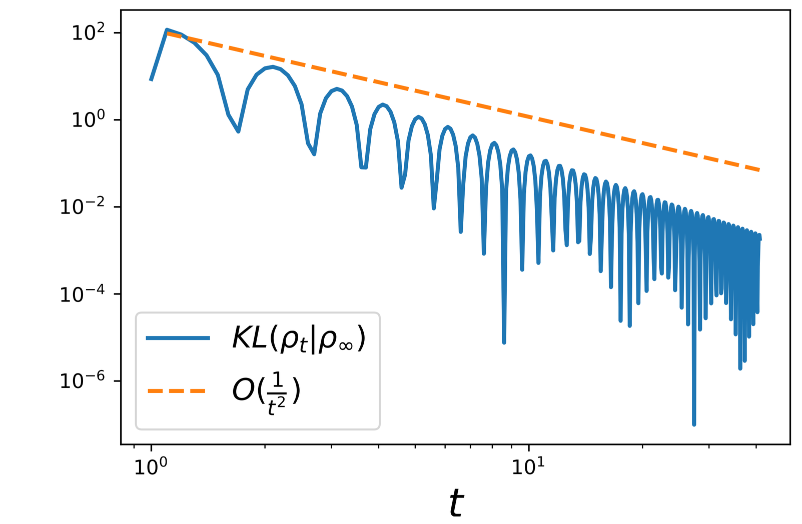
- (non-parametric) Diffusion-map approximation

$$\nabla\log(\rho(x)) \approx -\frac{1}{\epsilon}\frac{\sum_{i=1}^N k_\epsilon(x,X^i)(x-X^i)}{\sum_{i=1}^N k_\epsilon(x,X^i)}$$

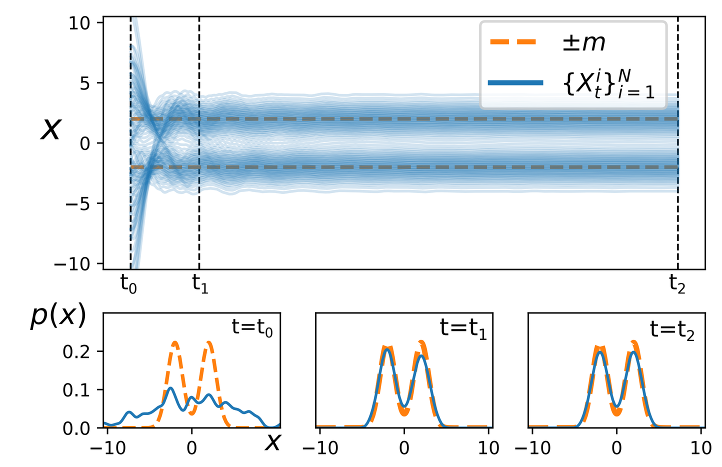
where $k_\epsilon(\cdot,\cdot)$ is the diffusion-map kernel

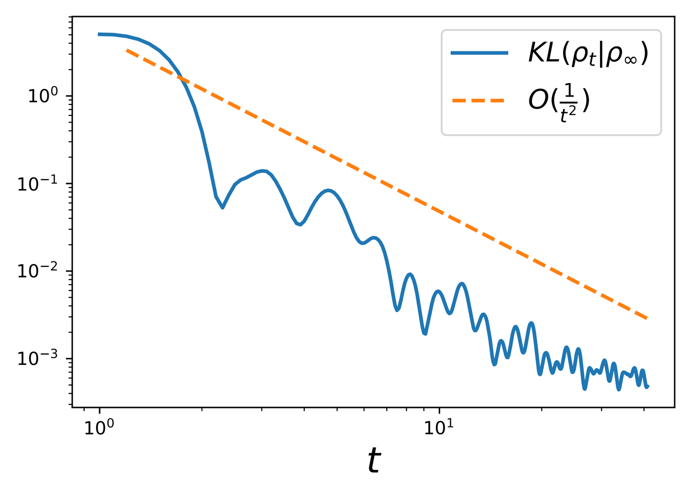
Numerical example: Target is Gaussian



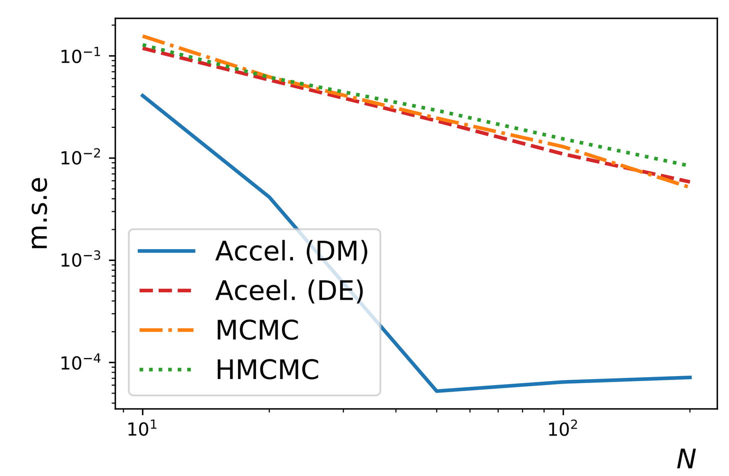


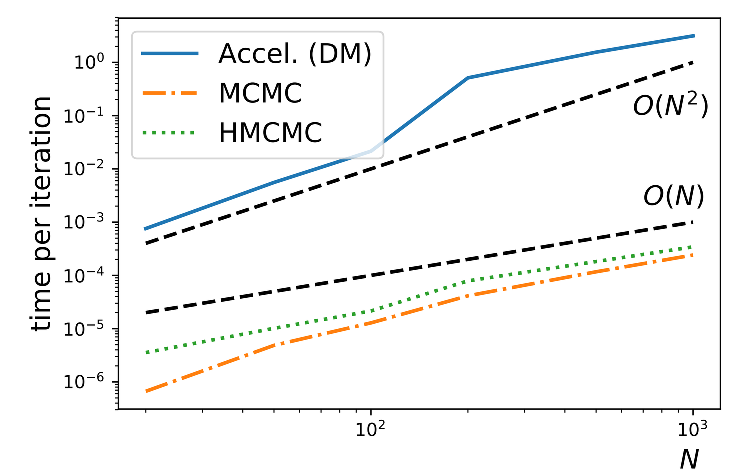
Numerical example: Target is mixture of Gaussians





Numerical comparison





Acknowledgement

Financial support from the NSF grant CMMI-1462773 and ARO grant W911NF1810334 is gratefully acknowledged.