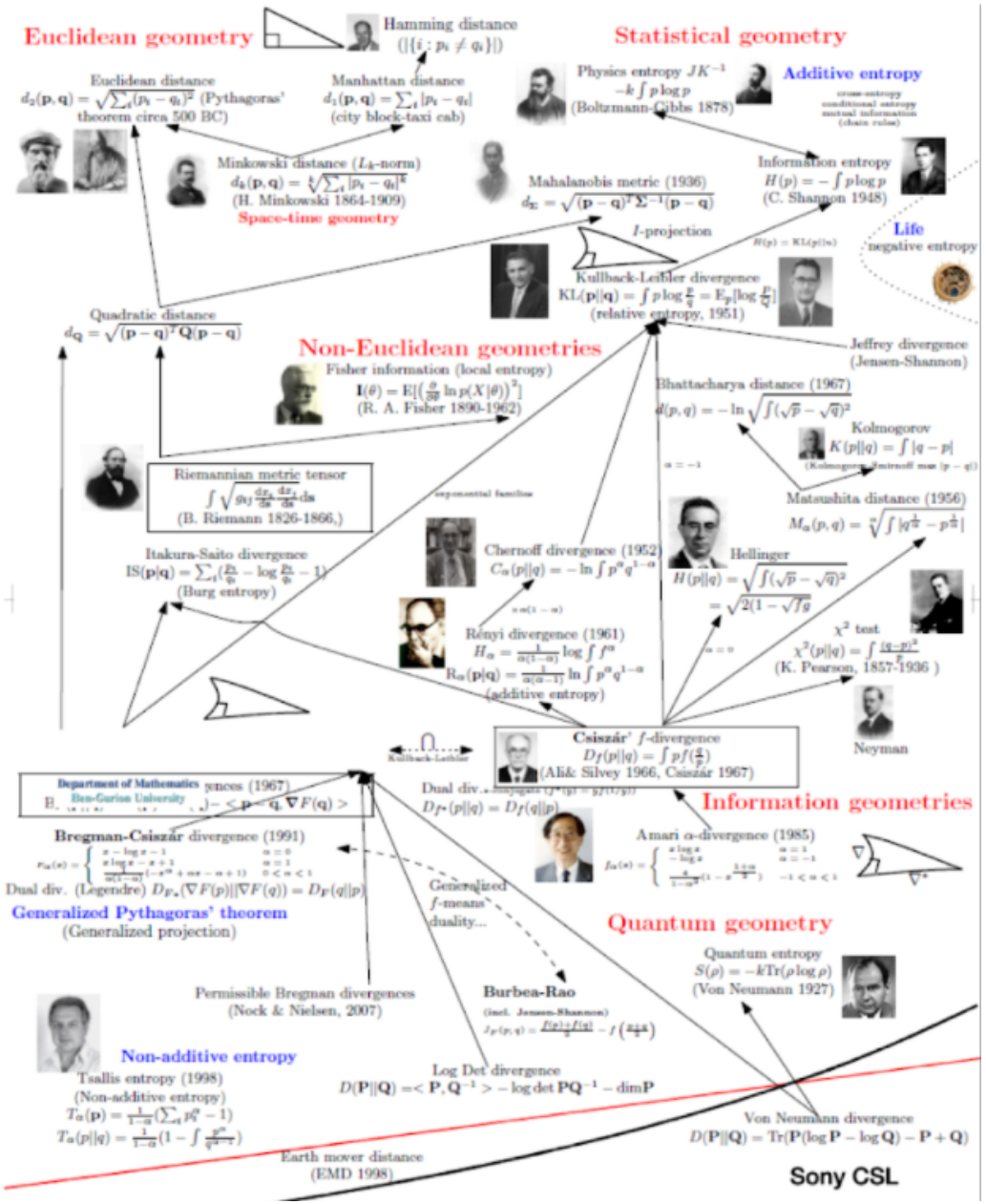
Transport information Newton's flow

Wuchen Li

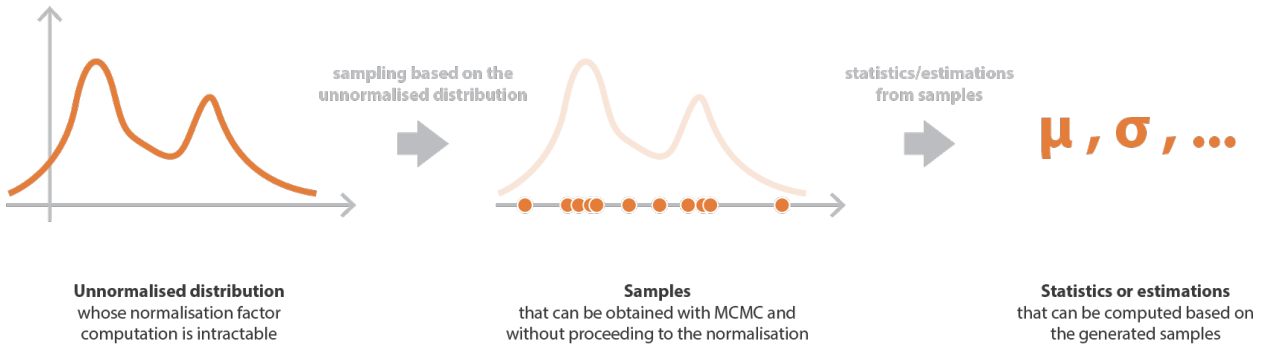
This is based on a joint work with Yifei Wang.

Divergence, metric and AI optimization



Bayesian inference

- A powerful tool in
  - Modeling complex data;
  - Quantifying uncertainty;
- Inverse problems, information science, physics and scientific computing;
- Main problem
  - Given a prior distribution, generate samples from posterior distributions;
  - Generate samples from an intractable distribution $\rho^*(x) \propto \exp(-f(x))$.



Langevin dynamics

- Consider the over-damped Langevin dynamics by

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t,$$

  where $X_t \in \mathbb{R}^d$ and B_t is the standard Brownian motion in $\mathbb{R}^d$.
- Denote $X_t \sim \rho_t$. The evolution of ρ_t satisfies the Fokker-Planck equation

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t.$$

- The invariant distribution satisfies

$$\rho^*(x) = \frac{1}{K} e^{-f(x)}, \qquad K = \int e^{-f(y)} dy.$$

Optimization in probability space

- Consider the optimization problem in probability space

$$\min_{\rho \in \mathcal{P}(\Omega)} E(\rho),$$

where

$$\mathcal{P}(\Omega) = \Big\{ \rho \in \mathcal{F}(\Omega) \colon \int_\Omega \rho dx = 1, \quad \rho \geq 0. \Big\}.$$

- A typical choice of objective functional in Bayesian inference problem is the KL divergence:

$$E(\rho) = \mathrm{D}_{\mathrm{KL}}(\rho \| e^{-f}) = \int \rho \log \frac{\rho}{e^{-f}} dx.$$

Information metrics and optimizations

Typical examples of information metrics include both Fisher-Rao metric (information geometry) and Wasserstein-2 metric (optimal transport).

Information geometry

- AI Inference problems: f–divergence, Amari α–divergence etc.
- AI optimization methods: Natural gradient (Amari); ADAM (Kingma et.al. 2014); Stochastic relaxation (Malago) and many more in book *Information geometry* (Ay et.al.)

Optimal transport

- AI and Machine learning: Wasserstein Training of Boltzmann Machines (Cuturi et.al. 2015); Learning from Wasserstein Loss (Frogner et.al. 2015); Wasserstein GAN (Bottou et.al. 2017); see NIPS, ICLR, ICML 2015– 2020;
- Gradient flows: (Jordan, Kinderlehrer, Otto, Villani, Slepcev, Carillo, et.al.);

Metrics in probability space

- Tangent space $T_\rho \mathcal{P}(\Omega) = \left\{\sigma \in \mathcal{F}(\Omega) : \int \sigma dx = 0.\right\}$;
- Cotangent space $T_\rho^* \mathcal{P}(\Omega)$: Equivalent to $\mathcal{F}(\Omega)/\mathbb{R}$;
- Metric tensor $\mathcal{G}(\rho) : T_\rho \mathcal{P}(\Omega) \to T_\rho^* \mathcal{P}(\Omega)$
- Metric: inner product in tangent space

$$g_\rho(\sigma_1, \sigma_2) = \int \Phi_1 \mathcal{G}(\rho)^{-1} \Phi_2 dx,$$

Here Φ_i is the solution to $\sigma_i = \mathcal{G}(\rho)^{-1}\Phi_i$, $i = 1, 2$.

Fisher-Rao metric

- Inverse of Fisher-Rao metric

$$\mathcal{G}^F(\rho)^{-1}\Phi = \rho(\Phi - \int \Phi\rho dx), \quad \Phi \in T^*_\rho \mathcal{P}(\Omega).$$

- Fisher-Rao metric: for $\sigma_1, \sigma_2 \in T_\rho \mathcal{P}(\Omega)$,

$$g^F_\rho(\sigma_1, \sigma_2) = \int \rho\Big(\Phi_1 - \int \Phi_1(y)\rho(y)dy, \Phi_2 - \Phi_2(y)\rho(y)dy\Big)dx,$$

where Φ_i is the solution to $\sigma_i = \rho(\Phi_i - \int \Phi_i \rho dx)$, $i = 1, 2$.

Wasserstein metric

- Inverse of Wasserstein metric tensor

$$\mathcal{G}^W(\rho)^{-1}\Phi = -\nabla \cdot (\rho \nabla \Phi), \quad \Phi \in T^*_\rho \mathcal{P}(\Omega).$$

- Wasserstein metric: for $\sigma_1, \sigma_2 \in T_\rho \mathcal{P}(\Omega)$,

$$g^W_\rho(\sigma_1, \sigma_2) = \int \rho \left\langle \nabla \Phi_1, \nabla \Phi_2 \right\rangle dx,$$

where Φ_i is the solution to $\sigma_i = -\nabla \cdot (\rho \nabla \Phi_i)$, $i = 1, 2$.

Gradient flows

- Gradient flow for $E(\rho)$ in $(\mathcal{P}(\Omega), \mathcal{G}(\rho))$

$$\partial_t \rho_t = -\mathcal{G}(\rho_t)^{-1} \frac{\delta E(\rho_t)}{\delta \rho_t}.$$

- Example: Fisher-Rao gradient flow of KL forms L^2–Newton's method:

$$\begin{aligned}\partial_t \rho_t =& -\operatorname{grad}^F \mathrm{D}_{\mathrm{KL}}(\rho_t \| \rho^*) \\ =& -\rho(\log \rho + f - \int \rho(\log \rho + f) dx).\end{aligned}$$

- Example: Wasserstein gradient flow of KL divergence forms the Fokker-Planck equation:

$$\begin{aligned}\partial_t \rho_t =& -\operatorname{grad}^W \mathrm{D}_{\mathrm{KL}}(\rho_t \| \rho^*) \\ =& \nabla \cdot (\rho_t \nabla (f + \log \rho_t + 1)) \\ =& \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t.\end{aligned}$$

Different viewpoint of Langevin dynamics

- SDE
$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t.$$
- Lagrangian viewpoint (Particle formulation)
$$dX_t = -\nabla f(X_t)dt - \nabla \log \rho_t(X_t)dt.$$
- Eulerian viewpoint
$$\partial_t \rho = \nabla \cdot (\rho \nabla f) + \Delta \rho.$$

First-order algorithms in AI and inverse problems

Various sampling methods are first-order methods in probability space based on various metrics.

- Stein metric and Stein variational gradient descent[1]:

$$\dot{X}_t = \int (-K(X_t, y)\nabla_y f(y) + \nabla_x K(X_t, y))\rho(y)dy,$$

  where K is a given matrix kernel function.

- Wasserstein-Kalman metric and Ensemble Kalman sampling[2]:

$$\dot{X}_t = -\mathcal{C}(\rho_t)\nabla f(X_t) + \sqrt{2\mathcal{C}(\rho_t)}\dot{B}_t$$

  where $\mathcal{C}(\rho)$ is the covariance matrix operator.

[1] Liu and Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. NIPS, 2016.

[2] Garbuno-Inigo, Hoffmann, Li, and Stuart. Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler. SIAM applied dynamical system, 2019.

Related works

- KL divergence+Wasserstein metric + Gradient descent = Langevin dynamics (Optimal transport, Jordan, Kinderlehrer, Otto, Villani et.al.)
- KL divergence+ Fisher-Rao metric+Gradient descent=Birth-death dynamics (Information geometry, Amari, et. al.)
- KL divergence +Stein metric + Gradient descent = Stein variational gradient descent (Liu et.al.)
- KL divergence+ Wasserstein-Kalman metric + Gradient descent = Ensemble Kalman sampling (Inigo, Hoffman, Li, Stuart et.al.)

Divergence+Metric + Newton's method = ?

Goal: Transport Newton's method

Instead of working on first order method for Bayesian optimization problems, we propose to construct the **second-order** method in optimal transport to accelerate computations.

- What is the **Newton's method** in probability space? In particular, what is the Newton's flow of KL divergence functional?
- What is the "Newton's" Langevin dynamics?

Related works on second-order methods in AI

- Optimization problems on Riemannian manifold[1]
- Discrete probability simplex with Fisher-Rao metric and exponential family models[2]
- Second-order methods for the Stein variational gradient descent direction[3] [4]
- Newton-type MCMC method (HAMCMC)[5]

[1] S. Smith. Optimization techniques on Riemannian manifolds. Fields institute communications, 1994.

[2] L. Malagó and G. Pistone. Combinatorial optimization with information geometry: The newton method. Entropy, 2014.

[3] G. Detommaso, T. Cui, Y. Marzouk, A. Spantini, and R. Scheichl. A Stein variational Newton method. In Advances in Neural Information Processing Systems, 2018.

[4] P. Chen, K. Wu, J. Chen, T. Roseberry, and O. Ghattas. Projected Stein variational Newton: A fast and scalable bayesian inference method in high dimensions. NIPS, 2019.

[5] U. Simsekli, R. Badeau, T. Cemgil, and G. Richard. Stochastic quasi-Newton Langevin Monte Carlo. ICML, 2016.

Newton's flow

The Newton's flow follows:

$$\dot{\rho} = -\mathrm{Hess}E(\rho)^{-1}\mathrm{grad}E(\rho)$$

We need to understand the Hessian operator in density manifold.

Hessian operator

Consider a Taylor expansion:

$$E(\rho(t)) = E(\rho(0)) + t\frac{d}{dt}E(\rho(t))|_{t=0} + \frac{t^2}{2}\frac{d^2}{dt^2}E(\rho(t))|_{t=0} + o(t^2),$$

where $\rho(0) = \rho$, $\partial_t \rho = \sigma$ satisfies the geodesic equation in density manifold. In the case of L^2–Wasserstein metric, we have

$$\begin{aligned}
&\frac{d^2}{dt^2}E(\rho(t))|_{t=0} = \mathrm{Hess}_W E(\rho)(\sigma,\sigma) \\
&= \int\int \nabla_x \nabla_y \frac{\delta^2}{\delta\rho(x)\delta\rho(y)} E(\rho)(\nabla_x\Phi(x), \nabla_y\Phi(y))\rho(x)\rho(y)dxdy \\
&\quad + \int \nabla_x^2 \frac{\delta}{\delta\rho(x)} E(\rho)(\nabla_x\Phi(x), \nabla_x\Phi(x))\rho(x)dx,
\end{aligned}$$

where $\sigma = -\nabla\cdot(\rho\nabla\Phi)$.

Derivation of Newton's direction

Consider

$$\min_{\sigma \in T_\rho \mathcal{P}} \; g_\rho(\mathrm{grad} E(\rho), \sigma) + \frac{1}{2} \mathrm{Hess} E(\rho)(\sigma, \sigma)$$

In the case of Wasserstein metric, we consider $\sigma = -\nabla \cdot (\rho \nabla \Phi)$, where Φ satisfies the following variational problem:

$$\begin{aligned}
\min_{\Phi \in T_\rho^* \mathcal{P}} \quad & \int (\nabla \Phi(x), \nabla \frac{\delta}{\delta \rho(x)} E(\rho)) \rho(x) dx \\
& + \frac{1}{2} \int \int \nabla_x \nabla_y \frac{\delta^2}{\delta \rho(x) \delta \rho(y)} E(\rho) (\nabla_x \Phi(x), \nabla_y \Phi(y)) \rho(x) \rho(y) dx dy \\
& + \frac{1}{2} \int \nabla_x^2 \frac{\delta}{\delta \rho(x)} E(\rho) (\nabla \Phi(x), \nabla \Phi(x)) \rho(x) dx,
\end{aligned}$$

Wasserstein Newton's direction

The Newton's direction satisfies

$$\mathrm{Hess}_W E(\rho)\sigma = -\,\mathrm{grad}_W E(\rho).$$

Denote

$$\sigma = -\nabla\cdot(\rho\nabla\Phi^{\mathrm{Newton}}),$$

then Φ^{Newton} satisfies the following equation

$$\begin{aligned}
&\nabla_x\cdot\Big(\rho(x)\int\nabla_x\nabla_y\frac{\delta^2}{\delta\rho(x)\delta\rho(y)}E(\rho)\nabla_y\Phi(y)\rho(y)dy\Big)\\
&+\nabla_x\cdot\Big(\rho(x)\nabla_x^2\frac{\delta}{\delta\rho(x)}E(\rho)\nabla_x\Phi(x)\Big)\\
=&\nabla_x\cdot\Big(\rho(x)\nabla_x\frac{\delta}{\delta\rho(x)}E(\rho)\Big).
\end{aligned}$$

Hessian operators and Gamma calculus

If $E(\rho) = \mathrm{D}_{\mathrm{KL}}(\rho \| \rho^*)$, then we can further reformulate the Hessian term by

$$\begin{aligned}
&\mathrm{Hess}_W E(\rho)(\sigma, \sigma) \\
&= \int \Gamma_2(\Phi, \Phi)\rho(x)dx \\
&= \int \Big(\mathrm{tr}(\nabla^2\Phi, \nabla^2\Phi) + \nabla^2 f(\nabla\Phi, \nabla\Phi) \Big)(x)\rho(x)dx
\end{aligned}$$

- Wasserstein Hessian operator (Otto, Villani et.al.)=Weak form of Gamma calculus (Bakry-Emery et.al.)
- Fisher-Rao +Wasserstein Christoffel symbol=Weak form of Gamma calculus (Li, 2018).
- Extended in sub-Riemannian density manifold=Weak form of Gamma z calculus (Feng and Li, 2019).

Wasserstein Newton's flow of KL divergence

Theorem

For a density $\rho^*(x) \propto \exp(-f(x))$, where f is a given function, denote the KL divergence between ρ and ρ^* by

$$\mathrm{D}_{KL}(\rho \| \rho^*) = \int \rho \log \frac{\rho}{e^{-f}} dx - \log Z,$$

where $Z = \int \exp(-f(x))dx$. Then the Wasserstein Newton's flow of KL divergence follows

$$\begin{cases} \partial_t \rho_t + \nabla \cdot (\rho_t \nabla \Phi_t) = 0 \\ \nabla^2 : (\rho_t \nabla^2 \Phi_t) - \nabla \cdot (\rho_t \nabla^2 f \nabla \Phi_t) - \nabla \cdot (\rho_t \nabla f) - \Delta \rho_t = 0. \end{cases}$$

Newton's Langevin dynamics

Theorem

Consider the Newton's Langevin dynamics

$$dX_t = \nabla \Phi_t^{\mathrm{Newton}}(X_t)dt,$$

where $\Phi_t^{\mathrm{Newton}}(x)$ *follows Wasserstein Newton's direction equation:*

$$\nabla^2 : (\rho_t \nabla^2 \Phi_t) - \nabla \cdot (\rho_t \nabla^2 f \nabla \Phi_t) - \nabla \cdot (\rho_t \nabla f) - \Delta \rho_t = 0.$$

Here X_0 *follows an initial distribution* ρ^0 *and* ρ_t *is the distribution of* X_t*. Then,* ρ_t *satisfies Wasserstein Newton's flow with an initial value* $\rho_0 = \rho^0$*.*

$$\text{Gradient flow} \Longrightarrow \text{Newton's flow}$$

Langevin dynamics	$dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$	
	$\downarrow$	
Density formulation	$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t$	$\partial_t \rho_t = -\nabla \cdot (\rho_t \nabla \Phi_t^{\mathrm{Newton}})$
Particle formulation	$dX_t = -\nabla f(X_t)dt - \nabla \log \rho_t(X_t)dt$	$dX_t = \nabla \Phi_t^{\mathrm{Newton}}(X_t)dt$

Figure: The relation among Wasserstein gradient flow, Newton's flow and Langevin dynamics. Our approach derive the particle formulation of Wasserstein Newton's flow of KL divergence.

Wasserstein Newton's flows in Gaussian families

Proposition

Suppose that ρ_0, ρ^* are Gaussian distributions with zero means and their covariance matrices are Σ_0 and Σ^*. Suppose $E(\Sigma)$ evaluates the KL divergence from ρ to ρ^*:

$$E(\Sigma) = \frac{1}{2}\left(\mathrm{tr}(\Sigma(\Sigma^*)^{-1}) - n - \log\det\left(\Sigma(\Sigma^*)^{-1}\right)\right).$$

Let (Σ_t, S_t) satisfy

$$\begin{cases} \dot{\Sigma}_t - 2(S\Sigma_t + \Sigma S_t) = 0, \\ 2\Sigma_t S_t(\Sigma^*)^{-1} + 2(\Sigma^*)^{-1} S_t \Sigma_t + 4S_t = -(\Sigma_t(\Sigma^*)^{-1} + (\Sigma^*)^{-1}\Sigma_t - 2I). \end{cases}$$

with initial values $\Sigma_t|_{t=0} = \Sigma_0$ and $S_t|_{t=0} = 0$. Thus, for any $t \geq 0$, Σ_t is well-defined and stays positive definite.

Proposition (Newton's Langevin dynamics in 1D Gaussian families)

Assume that $f(x) = (2\Sigma^*)^{-1}(x-\mu^*)^2$, where $\Sigma^* > 0$ and μ^* are given. Suppose that the particle system X_0 follows the Gaussian distribution. Then X_t follows a Gaussian distribution with mean μ_t and variance Σ_t. The corresponding NLD satisfies

$$dX_t = \left(\frac{\Sigma^* - \Sigma}{\Sigma^* + \Sigma_t} X_t - \frac{2\Sigma^*}{\Sigma^* + \Sigma_t} \mu_t + \mu^* \right) dt.$$

And the evolution of μ_t and Σ_t satisfies

$$d\mu_t = (-\mu_t + \mu^*)dt, \quad d\Sigma_t = 2\frac{\Sigma^* - \Sigma_t}{\Sigma^* + \Sigma_t}\Sigma_t dt.$$

The explicit solutions of μ_t and Σ_t satisfy

$$\mu_t = e^{-t}(\mu_0 - \mu^*) + \mu^*, \quad \Sigma_t = \Sigma^* + (\Sigma_0 - \Sigma^*)e^{-t}\sqrt{\frac{e^{-2t}(\Sigma_0 - \Sigma^*)^2}{4\Sigma_0^2} + \frac{1}{\Sigma_0 \Sigma^*}}.$$

Information Newton's method

- General update rule of Information Newton's method

$$\rho_{k+1} = \mathrm{Exp}_{\rho_k}(\alpha_k \Phi_k), \quad \mathcal{H}_E(\rho_k)\Phi_k + \mathcal{G}(\rho_k)^{-1}\frac{\delta E}{\delta \rho_k} = 0,$$

where $\mathrm{Exp}_{\rho_k}(\cdot)$ is the exponential map at ρ_k.

Riemannian structure of probability space

- Define the distance $\mathcal{D}(\rho_0, \rho_1)$

$$\mathcal{D}(\rho_0, \rho_1)^2 = \inf_{\hat{\rho}_s, s\in[0,1]} \left\{ \int_0^1 \int \partial_s \hat{\rho}_s \mathcal{G}(\hat{\rho}_s)^{-1} \partial_s \hat{\rho}_s dx ds : \hat{\rho}_s|_{s=0} = \rho_0, \hat{\rho}_s|_{s=1} = \rho_1 \right\}.$$

- Denote the inner product on cotangent space $T^*_\rho \mathcal{P}(\Omega)$ by

$$\langle \Phi_1, \Phi_2 \rangle_\rho = \int \Phi_1 \mathcal{G}(\rho)^{-1} \Phi_2 dx, \quad \Phi_1, \Phi_2 \in T^*_\rho \mathcal{P}(\Omega),$$

and $\|\Phi\|_\rho^2 = \langle \Phi, \Phi \rangle_\rho$.

Parallelism and high-order derivative

Definition (Parallelism)

We say that $\tau : T^*_{\rho_0}\mathcal{P}(\Omega) \to T^*_{\rho_1}\mathcal{P}(\Omega)$ is a parallelism from ρ_0 to ρ_1, if for all $\Phi_1, \Phi_2 \in T_{\rho_0}\mathcal{P}(\Omega)$, it follows

$$\langle \Phi_1, \Phi_2 \rangle_{\rho_0} = \langle \tau\Phi_1, \tau\Phi_2 \rangle_{\rho_1} .$$

- $\nabla^n E(\rho)$ is a n-form on the cotangent space $T^*_\rho\mathcal{P}(\Omega)$

$$\nabla^n E(\rho)(\Phi_1, \ldots, \Phi_n) = \frac{\partial}{\partial s}\nabla^{n-1}E(\mathrm{Exp}_\rho(s\Phi_n))(\tau_s\Phi_1, \ldots, \tau_s\Phi_{n-1})\Big|_{s=0},$$

  where τ_s is the parallelism from ρ to $\mathrm{Exp}_\rho(s\Phi_n)$.

Assumptions

Assumption

Assume that there exists $\epsilon, \delta_1, \delta_2, \delta_3 > 0$, such that for all ρ satisfying $\mathcal{D}(\rho, \rho^*) < \epsilon$, it follows

$$\nabla^2 E(\rho)(\Phi_1, \Phi_1) \geq \delta_1 \|\Phi_1\|_\rho^2,$$

$$\nabla^2 E(\rho)(\Phi_1, \Phi_1) \leq \delta_2 \|\Phi_1\|_\rho^2,$$

$$|\nabla^3 E(\rho)(\Phi_1, \Phi_1, \Phi_2)| \leq \delta_3 \|\Phi_1\|_\rho^2 \|\Phi_2\|_\rho,$$

holds for all $\Phi_1, \Phi_2 \in T_\rho^* \mathcal{P}(\Omega)$.

Convergence analysis

Theorem

Suppose that the assumption holds, ρ_k satisfies $\mathcal{D}(\rho_k, \rho^*) < \epsilon$ and the step size $\tau_k = 1$. Then, we have

$$\mathcal{D}(\rho_{k+1}, \rho^*) = O(\mathcal{D}(\rho_k, \rho^*)^2).$$

Sketch of proof

- Denote $T_k = \mathrm{Exp}_{\rho_k}^{-1}(\rho^*)$.

Proposition

Suppose that the assumption holds. Let τ be the parallelism from ρ_k to ρ_{k+1}. There exists a unique $R_k \in T^*_{\rho_k}\mathcal{P}(\Omega)$ such that

$$T_k = \tau^{-1}T_{k+1} + \Phi_k + R_k.$$

Then, we have

$$\|T_{k+1}\|_{\rho_{k+1}} \leq \frac{\delta_3}{\delta_1}\|T_k\|^2_{\rho_k} + \frac{\delta_2}{\delta_1}\|R_k\|_{\rho_k}.$$

Sketch of proof

Lemma

For all $\Psi \in \mathcal{T}^*_{\rho_k}\mathcal{P}(\Omega)$, it follows

$$\int \Psi \mathcal{G}(\rho_k)^{-1} R_k dx = O(\|\Psi\|_{\rho_k} \|T_k\|^2_{\rho_k}).$$

- Taking $\Psi = R_k$ in Lemma yields $\|R_k\|_{\rho_k} = O(\|T_k\|^2_{\rho_k})$.
- Because the geodesic curve has constant speed, $\|T_k\|^2_{\rho_k} = \mathcal{D}(\rho_k, \rho^*)^2$.
- As a result, we have

$$\mathcal{D}(\rho_{k+1}, \rho^*) \leq \frac{\delta_2}{\delta_1}\mathcal{D}(\rho_k, \rho^*)^2 + \frac{\delta_3}{\delta_1}\|R_k\|_{\rho_k} = O(\mathcal{D}(\rho_k, \rho^*)^2).$$

Implementation in probability space

- The distribution $\{x_k^i\}_{i=1}^N$ follows $\rho_k(x)$.
- Update each particle by

$$x_{k+1}^i = x_k^i + \alpha_k \nabla \Phi_k(x_k^i), \quad i = 1, 2 \dots N.$$

  Φ_k is the solution to Wasserstein Newton's direction equation.
- How to compute Φ_k based on $\{x_k^i\}_{i=1}^N$?

Compute Wasserstein Newton's direction

Proposition

Suppose that $\mathcal{H} : T_\rho^* \mathcal{P}(\Omega) \to T_\rho \mathcal{P}$ is a linear self-adjoint operator and $\mathcal{H}$ is positive definite. Let $u \in T_\rho \mathcal{P}$. Then the minimizer of variational problem

$$\min_{\Phi \in T_\rho^* \mathcal{P}(\Omega)} J(\Phi) = \int (\Phi \mathcal{H} \Phi - 2u\Phi)\, dx,$$

satisfies $\mathcal{H}\Phi = u$, where $\Phi \in T_\rho^* \mathcal{P}(\Omega)$.

Approximation methods

- For strongly convex f. Equivalent to optimizing the variational problem:

$$\min_{\Phi \in T^*_{\rho_k}\mathcal{P}(\Omega)} J(\Phi) = \int \left(\|\nabla^2 \Phi\|_F^2 + \|\nabla \Phi\|^2_{\nabla^2 f} + 2 \langle \nabla f + \nabla \log \rho_k, \nabla \Phi \rangle \right) \rho_k dx.$$

  Here we denote $\|v\|_A^2 = v^T A v$.

- For general f, suppose that $\nabla^2 f(x) + \epsilon I$ is strictly convex for $x \in \Omega$. Consider a regularized problem

$$\min_{\Phi \in T^*_{\rho_k}\mathcal{P}(\Omega)} J(\Phi) = \int \left(\|\nabla^2 \Phi\|_F^2 + \|\nabla \Phi\|^2_{\nabla^2 f + \epsilon I} + 2 \langle \nabla f + \nabla \log \rho_k, \nabla \Phi \rangle \right) \rho_k dx.$$

Affine Wasserstein Newton's method

- $\Phi(x)$ takes the form $\Phi(x) = \frac{1}{2} x S x + b^T x$, where $S = \mathbf{diag}(s) \in \mathbb{R}^{n \times n}$ is a diagonal matrix.
- The variational problem turns to be

$$\min_{S \in \mathbb{S}^n, b \in \mathbb{R}^n} J(S, b) = \mathrm{tr}(S^2) + \frac{1}{N} \sum_{i=1}^{N} \left(\|S x_k^i + b\|^2_{\nabla^2 f(x_k^i) + \epsilon I} + 2 \left\langle S x_k^i + b, v_k^i \right\rangle \right).$$

Affine Wasserstein Newton's method

- Rewrite the objective function to be

$$
\begin{aligned}
&J(s,b)\\
=&\|s\|^2 + \frac{1}{N}\sum_{i=1}^{N} \mathrm{tr}((\mathbf{diag}(x_k^i)s+b)^T(\nabla^2 f(x_k^i)+\epsilon I)(\mathbf{diag}(x_k^i)s+b))\\
&+2\left\langle \mathbf{diag}(x_k^i)s+b, v_k^i\right\rangle\\
=&\begin{bmatrix} s\\ b\end{bmatrix}^T \mathbf{H}_k \begin{bmatrix} s\\ b\end{bmatrix} + 2\begin{bmatrix} s\\ b\end{bmatrix}^T u_k.
\end{aligned}
$$

where

$$
\mathbf{H}_k = \begin{bmatrix} I + \frac{1}{N}\sum_{i=1}^N \mathbf{diag}(x_k^i)(\nabla^2 f(x_k^i)+\epsilon I)\,\mathbf{diag}(x_k^i) & \frac{1}{N}\sum_{i=1}^N \mathbf{diag}(x_k^i)(\nabla^2 f(x_k^i)+\epsilon I)\\ \frac{1}{N}\sum_{i=1}^N (\nabla^2 f(x_k^i)+\epsilon I)\,\mathbf{diag}(x_k^i) & \frac{1}{N}\sum_{i=1}^N (\nabla^2 f(x_k^i)+\epsilon I)\end{bmatrix},
$$

$$
u_k = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^N \mathbf{diag}(x_k^i)v_k^i\\ \frac{1}{N}\sum_{i=1}^N v_k^i\end{bmatrix}.
$$

Algorithm 1 Affine Wasserstein Newton's method

Require: initial positions $\{x_0^i\}_{i=1}^N$, $\epsilon \geq 0$, step sizes α_k, maximum iteration K.

Set $k = 0$.

while $k < K$ and the convergence criterion is not met **do**

Compute $v_k^i = \nabla f(x_k^i) + \xi_k(x_k^i)$. Here ξ_k is an approximation of $\nabla \log \rho_k$.

Calculate $\mathbf{H}_k$ by

$$\mathbf{H}_k = \begin{bmatrix} I + \frac{1}{N}\sum_{i=1}^N \mathbf{diag}(x_k^i)(\nabla^2 f(x_k^i) + \epsilon I)\,\mathbf{diag}(x_k^i) & \frac{1}{N}\sum_{i=1}^N \mathbf{diag}(x_k^i)(\nabla^2 f(x_k^i) + \epsilon I) \\ \frac{1}{N}\sum_{i=1}^N (\nabla^2 f(x_k^i) + \epsilon I)\,\mathbf{diag}(x_k^i) & \frac{1}{N}\sum_{i=1}^N (\nabla^2 f(x_k^i) + \epsilon I) \end{bmatrix},$$

and formulate u_k by

$$u_k = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^N \mathbf{diag}(x_k^i) v_k^i \\ \frac{1}{N}\sum_{i=1}^N v_k^i \end{bmatrix}.$$

Compute s_k and b_k by

$$\begin{bmatrix} s_k \\ b_k \end{bmatrix} = -(\mathbf{H}_k)^{-1} u_k.$$

Update particle positions by

$$x_{k+1}^i = x_k^i + \alpha_k(\mathbf{diag}(s_k) x_k^i + b_k).$$

Set $k = k + 1$.

end while

Hybrid method

- Overdamped Langevin dynamics as gradient direction:

$$x_{k+1}^i = x_k^i + \alpha_k(S_k x_k^i + b_k) + \sqrt{2\lambda_k \alpha_k} B_k,$$

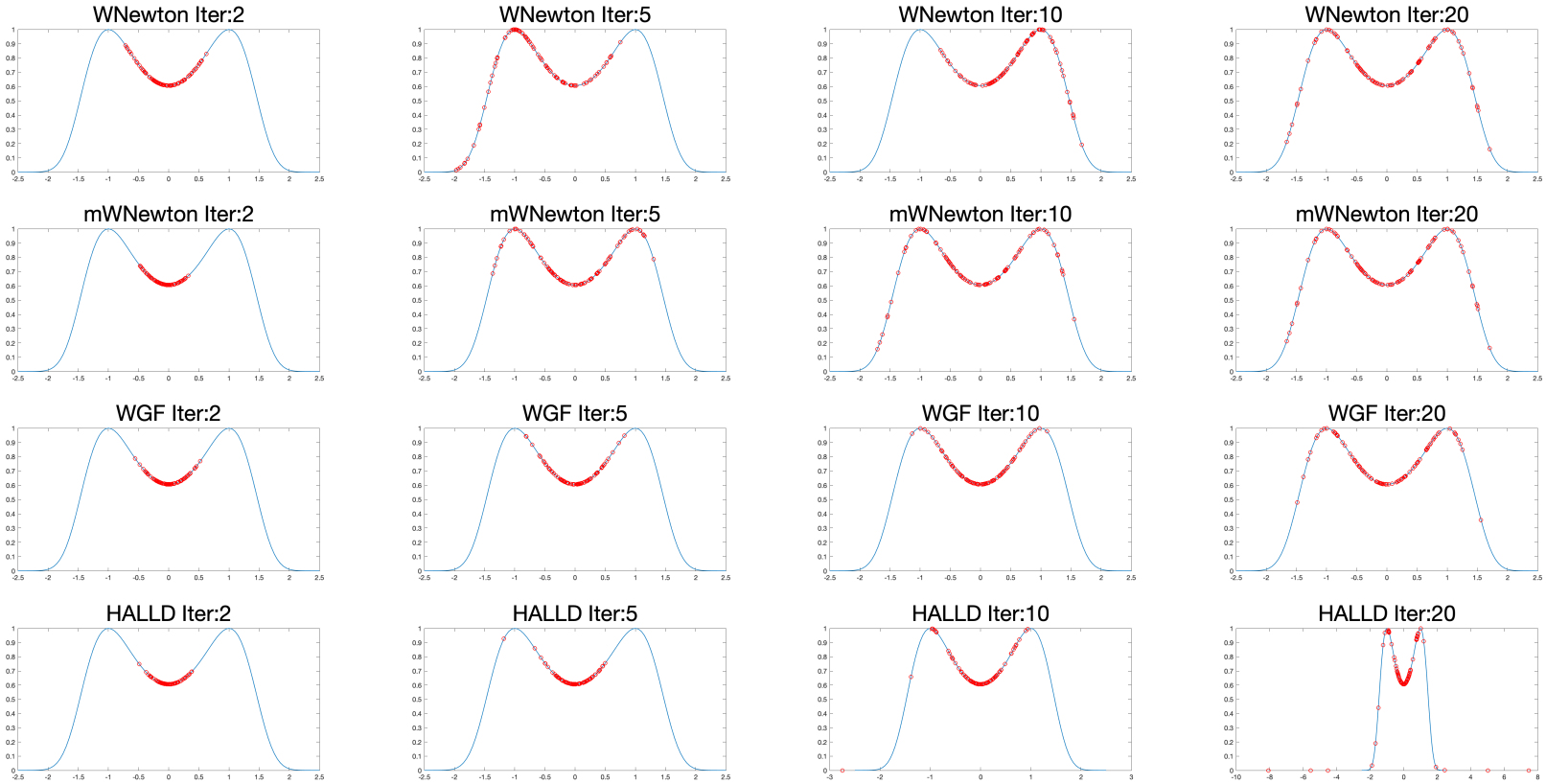
  where $B_k \sim \mathcal{N}(0, I)$.
- Lagrangian Langevin dynamics as gradient direction:

$$x_{k+1}^i = x_k^i + \alpha_k(S_k x_k^i + b_k - \lambda_k v_k^i).$$

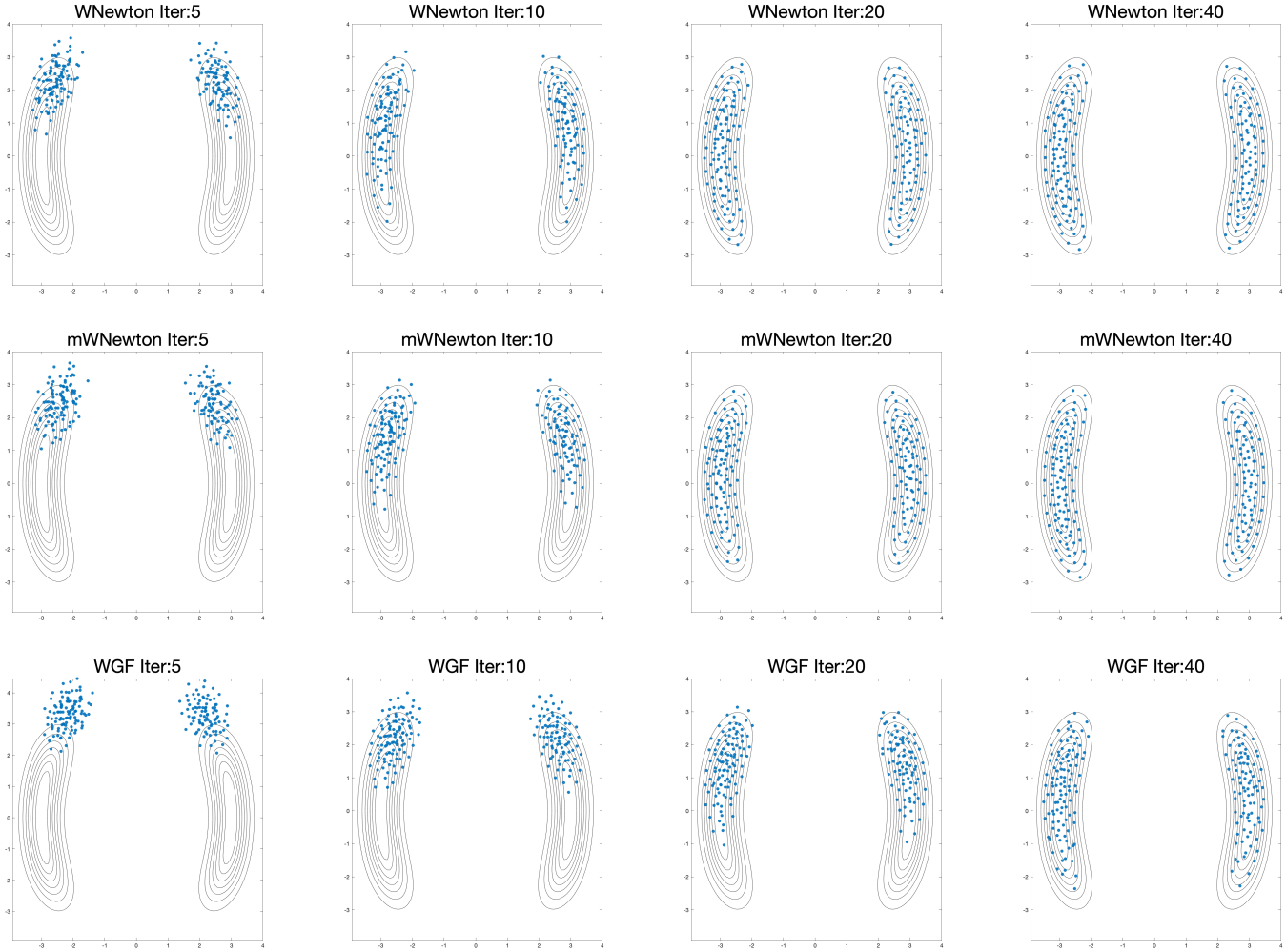
- Hybrid Langvien dynamics

$$dX_t = (\nabla \Phi_t^{\epsilon-\text{Newton}} - \lambda_t \nabla f)dt + \sqrt{2\lambda_t} dB_t,$$

1D Toy example



2D toy example



Gaussian families

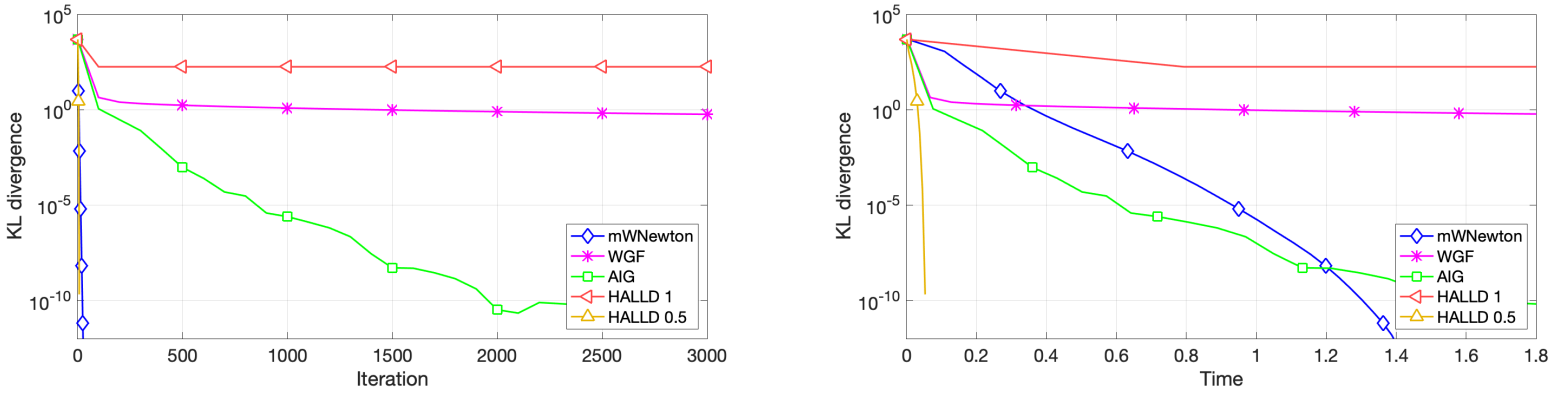


Figure: Comparison among WNewton, WGF, AIG and HALLD in Gaussian families. The conditional number $\kappa = 2 \times 10^4$. For Newton and HALLD 0.5, the markers are marked for every 5 iterations.

Bayesian logistic regression

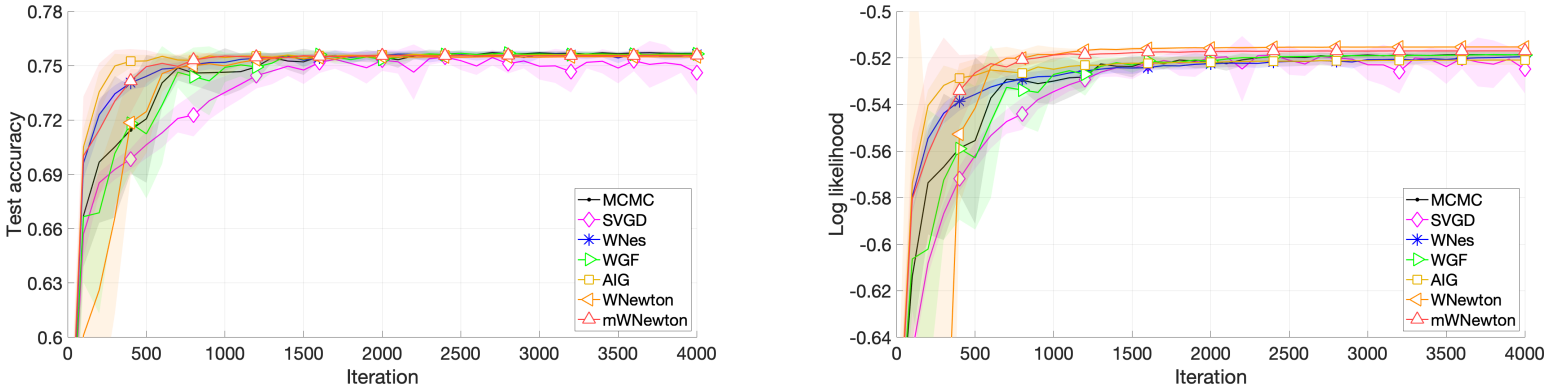


Figure: Comparison of different methods on Bayesian logistic regression, averaged over 10 independent trials. The shaded areas show the variance over 10 trials. Left: Test accuracy; Right: Test log-likelihood.

Discussion

- Design high-order optimization methods for Bayesian sampling, machine learning and inverse problems;
- Analysis on high-order derivatives from information metrics;
- Sampling efficient Quasi-Newton's method for information metrics;
- Other efficient method to approximate the Wasserstein Newton's direction;
- Newton's random walk;
- Information Newton's flow in probability models.