Kernelized Wasserstein Natural Gradient

Michael Arbel [1] Arthur Gretton [1] Wuchen Li [2] Guido Montúfar [3]

[1]Gatsby Computational Neuroscience Unit, UCL, London

[2]University of California, Los Angeles

[3]Max Planck Institute for Mathematics in the Sciences, Leipzig

November 20, 2019

Outline

- General problem and Motivation
- Wasserstein Natural Gradient
- Kernelized Wasserstein Natural Gradient
- Experiments

Motivation: General problem

Given a model ρ_θ of the form:

$$X \sim \rho_\theta \iff X = h_\theta(Z), \qquad Z \sim \nu$$

Consider the learning problem:

Motivation: General problem

Given a model ρ_θ of the form:

$$X \sim \rho_\theta \iff X = h_\theta(Z), \qquad Z \sim \nu$$

Consider the learning problem:

$$\min_\theta \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(Z))\, \mathrm{d}\nu(Z)$$

Motivation: General problem

Given a model ρ_θ of the form:

$$X \sim \rho_\theta \iff X = h_\theta(Z), \qquad Z \sim \nu$$

Consider the learning problem:

$$\min_\theta \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(\boxed{Z}))\, \mathrm{d}\nu(Z)$$

Latent variable

Motivation: General problem

Given a model ρ_θ of the form:

$$X \sim \rho_\theta \iff X = h_\theta(Z), \qquad Z \sim \nu$$

Image

Consider the learning problem:

$$\min_\theta \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(Z))\, \mathrm{d}\nu(Z)$$

Latent variable

Motivation: General problem

Given a model ρ_θ of the form:

$$X \sim \rho_\theta \iff X = h_\theta(Z), \qquad (Z, Y) \sim \nu$$

Consider the learning problem:

$$\min_\theta \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(Z), Y) \, \mathrm{d}\nu(Z, Y)$$

Motivation: General problem

Given a model ρ_θ of the form:

$$X \sim \rho_\theta \iff X = h_\theta(Z), \qquad (Z, Y) \sim \nu$$

Consider the learning problem:

$$\min_\theta \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(Z), Y)\, \mathrm{d}\nu(Z, Y)$$

Image

Motivation: General problem

Given a model ρ_θ of the form:

$$X \sim \rho_\theta \iff X = h_\theta(Z), \qquad (Z, Y) \sim \nu$$

Consider the learning problem:

$$\min_\theta \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(Z), Y)\, \mathrm{d}\nu(Z, Y)$$

Label

Image

Motivation: General problem

Given a model ρ_θ of the form:

$$X \sim \rho_\theta \iff X = h_\theta(Z), \qquad (Z, Y) \sim \nu$$

Logit

Consider the learning problem:

$$\min_\theta \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(Z), Y)\, \mathrm{d}\nu(Z, Y)$$

Image

Label

First order methods

Large scale models $\Rightarrow$ SGD

$$\theta_{t+1} = \theta_t - \gamma \widehat{\nabla \mathcal{L}}(\rho_{\theta_t})$$

First order methods

Large scale models $\Rightarrow$ SGD

$$\theta_{t+1} = \theta_t - \gamma \boxed{\widehat{\nabla \mathcal{L}}(\rho_{\theta_t})}$$

Euclidean gradient

First order methods

Large scale models $\Rightarrow$ SGD

$$\theta_{t+1} = \theta_t - \gamma \widehat{\nabla \mathcal{L}}(\rho_{\theta_t})$$

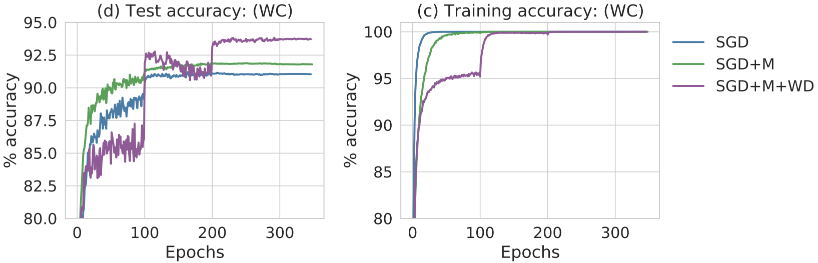
- Fast training (using autodiff and GPUs)
- Often gives impressive results

First order methods

Large scale models $\Rightarrow$ SGD

$$\theta_{t+1} = \theta_t - \gamma \widehat{\nabla \mathcal{L}}(\rho_{\theta_t})$$

- Fast training (using autodiff and GPUs)
- Often gives impressive results

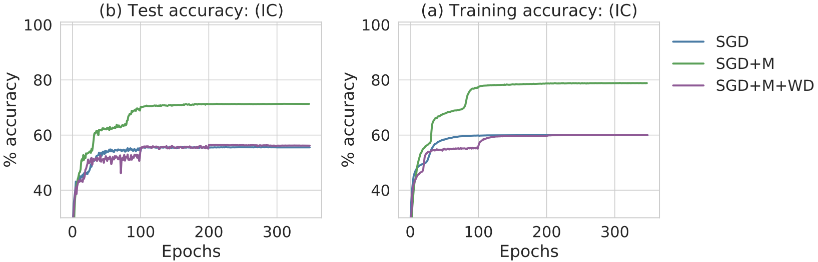


First order methods: Challenges

Large scale models $\Rightarrow$ SGD

$$\theta_{t+1} = \theta_t - \gamma \widehat{\nabla \mathcal{L}}(\rho_{\theta_t})$$

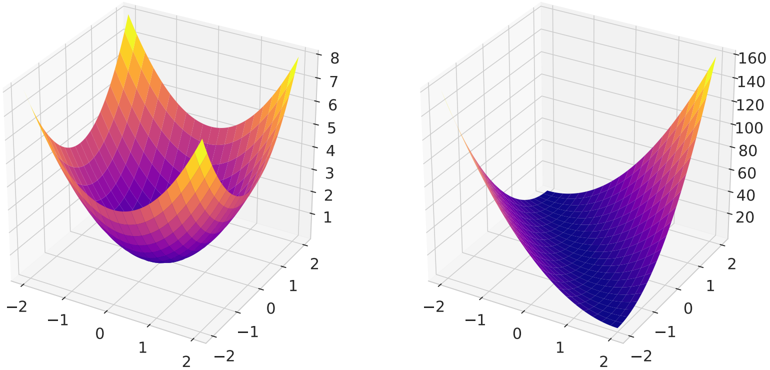
- Sensitive to parametrization
- Can fail miserably when the problem is ill-conditioned



Ill-conditioned problem

Definition (Ill-conditioned problem)

A problem $\min_\theta \mathcal{L}(\rho_\theta)$ is ill-conditioned if the hessian $H\mathcal{L}(\rho_\theta)$ at a local optimum θ^* has a high condition number: $\kappa := \frac{\lambda_{max}}{\lambda_{min}}$



Second order methods

Ill-conditioned problem $\Rightarrow$ Second order methods!!

$$\theta_{t+1} = \theta_t - \gamma \hat{G}(\theta_t)^{-1} \widehat{\nabla \mathcal{L}}(\rho_{\theta_t})$$

Second order methods

Ill-conditioned problem $\Rightarrow$ Second order methods!!

$$\theta_{t+1} = \theta_t - \gamma \hat{G}(\theta_t)^{-1} \widehat{\nabla \mathcal{L}(\rho_{\theta_t})}$$

Euclidean gradient

Second order methods

Ill-conditioned problem $\Rightarrow$ Second order methods!!

$$\theta_{t+1} = \theta_t - \gamma \boxed{\hat{G}(\theta_t)^{-1}} \boxed{\widehat{\nabla \mathcal{L}}(\rho_{\theta_t})}$$

Preconditioner

Euclidean gradient

Second order methods

Ill-conditioned problem $\Rightarrow$ Second order methods!!

$$\theta_{t+1} = \theta_t - \gamma \hat{G}(\theta_t)^{-1} \widehat{\nabla \mathcal{L}}(\rho_{\theta_t})$$

Second order information

Euclidean gradient

Second order methods

Ill-conditioned problem $\Rightarrow$ Second order methods!!

$$\theta_{t+1} = \theta_t - \gamma \hat{G}(\theta_t)^{-1} \widehat{\nabla \mathcal{L}}(\rho_{\theta_t})$$

Second order information

Euclidean gradient

- When the density of ρ_θ is available ...

Second order methods

Ill-conditioned problem $\Rightarrow$ Second order methods!!

$$\theta_{t+1} = \theta_t - \gamma \underbrace{\hat{G}(\theta_t)^{-1}}_{\text{Second order information}} \underbrace{\widehat{\nabla \mathcal{L}}(\rho_{\theta_t})}_{\text{Euclidean gradient}}$$

- When the density of ρ_θ is available ...
- Can choose $\hat{G}(\theta_t)$ as an estimator of the fisher information matrix:

$$G_F(\theta_t) = \int \nabla \rho_{\theta_t}(x) \nabla \rho_{\theta_t}(x)^\top \rho_\theta(x) \, \mathrm{d}x$$

$$\nabla^F \mathcal{L}(\rho_\theta) := G_F(\theta_t)^{-1} \nabla \mathcal{L}(\rho_{\theta_t})$$

Second order methods

Ill-conditioned problem $\Rightarrow$ Second order methods!!

$$\theta_{t+1} = \theta_t - \gamma \boxed{\hat{G}(\theta_t)^{-1}} \boxed{\widehat{\nabla \mathcal{L}}(\rho_{\theta_t})}$$

Second order information

Euclidean gradient

- When the density of ρ_θ is available ...
- Can choose $\hat{G}(\theta_t)$ as an estimator of the fisher information matrix:

$$G_F(\theta_t) = \int \nabla \rho_{\theta_t}(x) \nabla \rho_{\theta_t}(x)^\top \rho_\theta(x) \, \mathrm{d}x$$

$$\nabla^F \mathcal{L}(\rho_\theta) := G_F(\theta_t)^{-1} \nabla \mathcal{L}(\rho_{\theta_t})$$

- $\nabla^F \mathcal{L}(\rho_\theta)$ called the *Fisher Natural gradient*.

Second order methods

Ill-conditioned problem $\Rightarrow$ Second order methods!!

$$\theta_{t+1} = \theta_t - \gamma \underbrace{\hat{G}(\theta_t)^{-1}}_{\text{Second order information}} \underbrace{\widehat{\nabla \mathcal{L}}(\rho_{\theta_t})}_{\text{Euclidean gradient}}$$

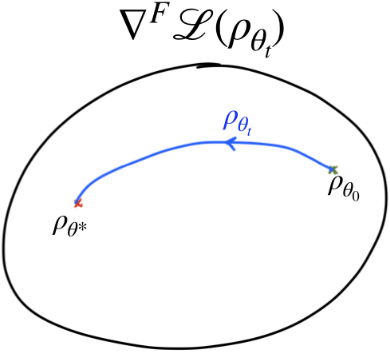
- When the density of ρ_θ is available ...
- Can choose $\hat{G}(\theta_t)$ as an estimator of the fisher information matrix:

$$G_F(\theta_t) = \int \nabla \rho_{\theta_t}(x) \nabla \rho_{\theta_t}(x)^\top \rho_\theta(x) \,\mathrm{d}x$$

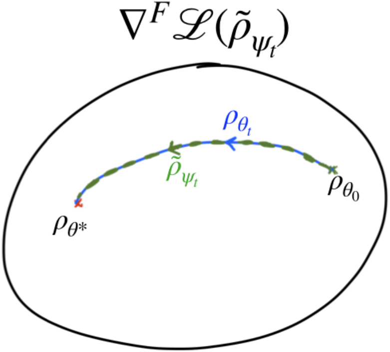
$$\nabla^F \mathcal{L}(\rho_\theta) := G_F(\theta_t)^{-1} \nabla \mathcal{L}(\rho_{\theta_t})$$

- $\nabla^F \mathcal{L}(\rho_\theta)$ called the *Fisher Natural gradient*.
- Robust to parametrization.

Second order methods: Robustness to parametrization



Second order methods: Robustness to parametrization



Second order methods: Robustness to parametrization

Fisher Natural gradient descent: $\theta_{t+1} = \theta_t - \gamma \nabla^F \mathcal{L}(\rho_{\theta_t})$

- Have a change of variables $\psi = \Psi(\theta)$ and write $\tilde{\rho}_\psi = \rho_\theta$.

Second order methods: Robustness to parametrization

Fisher Natural gradient descent: $\theta_{t+1} = \theta_t - \gamma \nabla^F \mathcal{L}(\rho_{\theta_t})$

- Have a change of variables $\psi = \Psi(\theta)$ and write $\tilde{\rho}_\psi = \rho_\theta$.
- Continuous-time limit of the Fisher natural gradient:

$$\dot{\theta}_t = -\nabla^F \mathcal{L}(\rho_{\theta_t}), \quad \theta_0$$
$$\dot{\psi}_t = -\nabla^F \mathcal{L}(\tilde{\rho}_{\psi_t}), \quad \psi_0 = \Psi(\theta_0)$$

Second order methods: Robustness to parametrization

Fisher Natural gradient descent: $\theta_{t+1} = \theta_t - \gamma \nabla^F \mathcal{L}(\rho_{\theta_t})$

- Have a change of variables $\psi = \Psi(\theta)$ and write $\tilde{\rho}_\psi = \rho_\theta$.
- Continuous-time limit of the Fisher natural gradient:

$$\dot{\theta}_t = -\nabla^F \mathcal{L}(\rho_{\theta_t}), \quad \theta_0$$
$$\dot{\psi}_t = -\nabla^F \mathcal{L}(\tilde{\rho}_{\psi_t}), \quad \psi_0 = \Psi(\theta_0)$$

- Robustness to parametrization $\Rightarrow \psi_t = \Psi(\theta_t)$

Second order methods: Robustness to parametrization

Fisher Natural gradient descent: $\theta_{t+1} = \theta_t - \gamma \nabla^F \mathcal{L}(\rho_{\theta_t})$

- Have a change of variables $\psi = \Psi(\theta)$ and write $\tilde{\rho}_\psi = \rho_\theta$.
- Continuous-time limit of the Fisher natural gradient:

$$\begin{aligned}\dot{\theta}_t &= -\nabla^F \mathcal{L}(\rho_{\theta_t}), \quad \theta_0 \\ \dot{\psi}_t &= -\nabla^F \mathcal{L}(\tilde{\rho}_{\psi_t}), \quad \psi_0 = \Psi(\theta_0)\end{aligned}$$

- Robustness to parametrization $\Rightarrow \psi_t = \Psi(\theta_t)$
- Doesn't hold for the euclidean gradient in general!

Second order methods: Challenges

Fisher Natural gradient descent: $\theta_{t+1} = \theta_t - \gamma \hat{G}_F(\theta_t)^{-1} \widehat{\nabla \mathcal{L}}(\rho_{\theta_t})$

[1] [Grosse and Martens, 2016, George et al., 2018]

Second order methods: Challenges

Fisher Natural gradient descent: $\theta_{t+1} = \theta_t - \gamma \hat{G}_F(\theta_t)^{-1} \widehat{\nabla \mathcal{L}}(\rho_{\theta_t})$

- Computational: Expensive to store and invert $\hat{G}_F(\theta)$ at every iteration .
- Prior works proposed cheap approximations of $\hat{G}_F(\theta)$ [1].
- Requires to know the density ρ_θ.

[1] [Grosse and Martens, 2016, George et al., 2018]

Contributions

A second order method based on the Wasserstein natural gradient which is:

- Robust to parametrization
- Doesn't require access to the density of the model
- Trades-off between accuracy and computational cost
- Comes with convergence rates.

General recipe for natural gradients

1. Choose a distance/divergence d defined on the model ρ_θ:

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + d(\rho_{\theta_t+u}, \rho_{\theta_t})$$

2. Second order expansion of d at ρ_θ:

$$d(\rho_{\theta_t+u}, \rho_{\theta_t}) \simeq \frac{1}{2} u^\top G(\theta_t) u$$

3. Solve:

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2} u^\top G(\theta_t) u$$

Recipe for Fisher natural gradients

1. Choose the *KL* as a divergence

$$\min_u \nabla\mathcal{L}(\rho_{\theta_t})^\top u + KL(\rho_{\theta_t+u}|\rho_{\theta_t})$$

2. Second order expansion of *KL* at ρ_θ:

$$KL(\rho_{\theta_t+u}, \rho_{\theta_t}) \simeq \frac{1}{2}u^\top G_F(\theta_t)u$$

3. Solve:

$$\min_u \nabla\mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2}u^\top G_F(\theta_t)u$$

Wasserstein Natural Gradient: Step 1

1. Choose $d(\rho_\theta, \rho_{\theta+u}) = \frac{1}{2} W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2\lambda} W_2^2(\rho_{\theta_t+u}, \rho_{\theta_t})$$

- Well defined even when ρ_θ doesn't admit a density

Wasserstein Natural Gradient: Step 1

1. Choose $d(\rho_\theta, \rho_{\theta+u}) = \frac{1}{2} W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2\lambda} W_2^2(\rho_{\theta_t+u}, \rho_{\theta_t})$$

- Well defined even when ρ_θ doesn't admit a density

$$W_2^2(\delta_{\theta_1}, \delta_{\theta_2}) = \|\theta_1 - \theta_2\|^2$$

Wasserstein Natural Gradient: Step 1

1. Choose $d(\rho_\theta, \rho_{\theta+u}) = \frac{1}{2} W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2\lambda} W_2^2(\rho_{\theta_t+u}, \rho_{\theta_t})$$

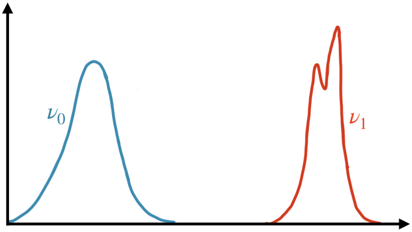
- W_2 has a nice geometric interpretation:

Wasserstein Natural Gradient: Step 1

1. Choose $d(\rho_\theta, \rho_{\theta+u}) = \frac{1}{2} W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2\lambda} W_2^2(\rho_{\theta_t+u}, \rho_{\theta_t})$$

- W_2 has a nice geometric interpretation:

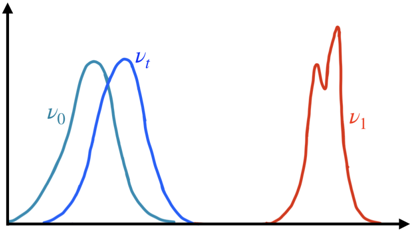


Wasserstein Natural Gradient: Step 1

1. Choose $d(\rho_\theta, \rho_{\theta+u}) = \frac{1}{2} W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2\lambda} W_2^2(\rho_{\theta_t+u}, \rho_{\theta_t})$$

- W_2 has a nice geometric interpretation:

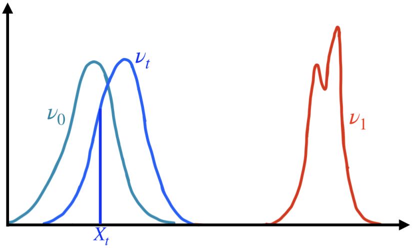


Wasserstein Natural Gradient: Step 1

1. Choose $d(\rho_\theta, \rho_{\theta+u}) = \frac{1}{2} W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2\lambda} W_2^2(\rho_{\theta_t+u}, \rho_{\theta_t})$$

- W_2 has a nice geometric interpretation:

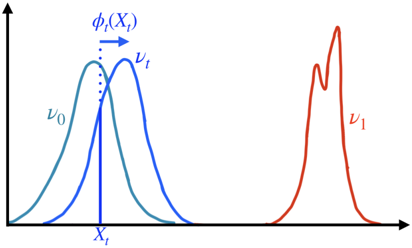


Wasserstein Natural Gradient: Step 1

1. Choose $d(\rho_\theta, \rho_{\theta+u}) = \frac{1}{2} W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2\lambda} W_2^2(\rho_{\theta_t+u}, \rho_{\theta_t})$$

- W_2 has a nice geometric interpretation:

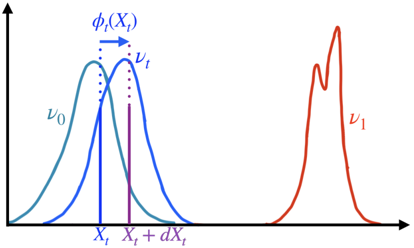


Wasserstein Natural Gradient: Step 1

1. Choose $d(\rho_\theta, \rho_{\theta+u}) = \frac{1}{2} W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2\lambda} W_2^2(\rho_{\theta_t+u}, \rho_{\theta_t})$$

- W_2 has a nice geometric interpretation:

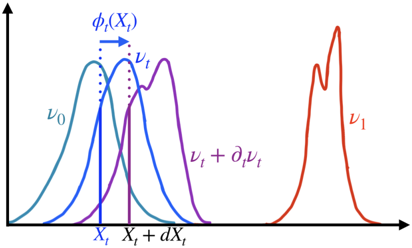


Wasserstein Natural Gradient: Step 1

1. Choose $d(\rho_\theta, \rho_{\theta+u}) = \frac{1}{2} W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2\lambda} W_2^2(\rho_{\theta_t+u}, \rho_{\theta_t})$$

- W_2 has a nice geometric interpretation:



Wasserstein Natural Gradient: Step 1

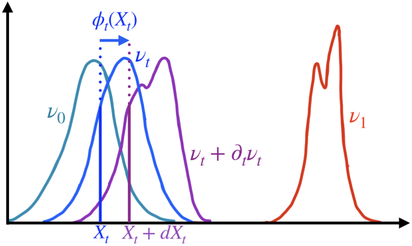
1. Choose $d(\rho_\theta, \rho_{\theta+u}) = \frac{1}{2} W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2\lambda} W_2^2(\rho_{\theta_t+u}, \rho_{\theta_t})$$

- W_2 has a nice geometric interpretation:

Continuity equation: $\partial_t \nu_t + div(\nu_t \phi_t) = 0$

Boundary conditions: $\nu_0 = p, \quad \nu_1 = q$

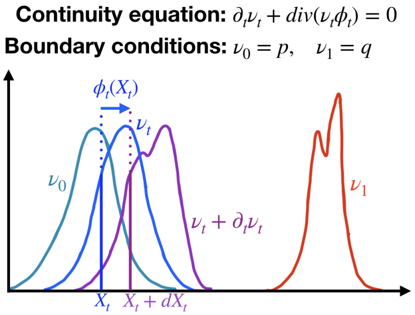


Wasserstein Natural Gradient: Step 1

1. Choose $d(\rho_\theta, \rho_{\theta+u}) = \frac{1}{2} W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2\lambda} W_2^2(\rho_{\theta_t+u}, \rho_{\theta_t})$$

- W_2 has a nice geometric interpretation:



Benamou-Brenier formula[2]: $W_2^2(p, q) := \inf_{(\nu_t, \phi_t)} \int_0^1 \int \|\phi_t(x)\|^2 \, \mathrm{d}\nu_t(x)$

[2] [Benamou and Brenier, 2000]

General recipe for natural gradients

1. Choose a distance/divergence d defined on the model ρ_θ:

$$\min_u \nabla\mathcal{L}(\rho_{\theta_t})^\top u + d(\rho_{\theta_t+u}, \rho_{\theta_t})$$

2. Second order expansion of d at ρ_θ:

$$d(\rho_{\theta_t+u}, \rho_{\theta_t}) \simeq \frac{1}{2} u^\top G(\theta_t) u$$

3. Solve:

$$\min_u \nabla\mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2} u^\top G(\theta_t) u$$

Wasserstein Natural Gradient: Step 2

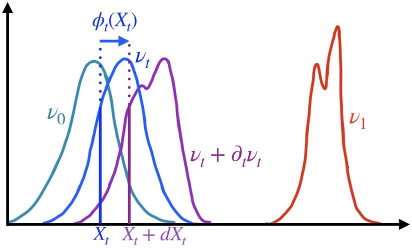
2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

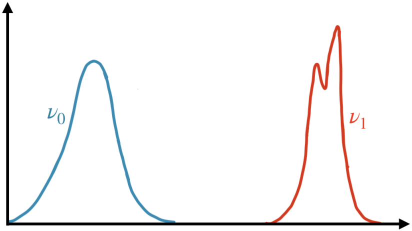
Continuity equation: $\partial_t \nu_t + div(\nu_t \phi_t) = 0$

Boundary conditions: $\nu_0 = p, \quad \nu_1 = q$



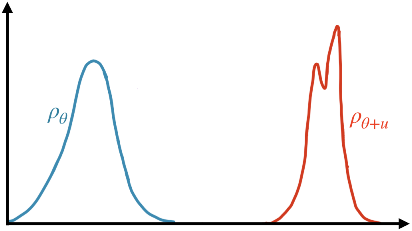
Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$



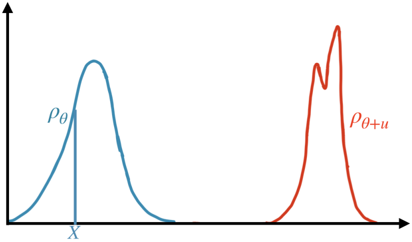
Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$



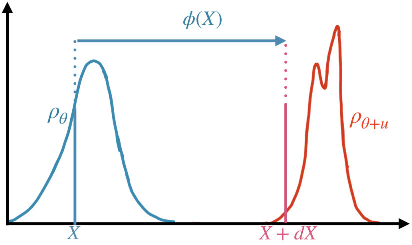
Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$



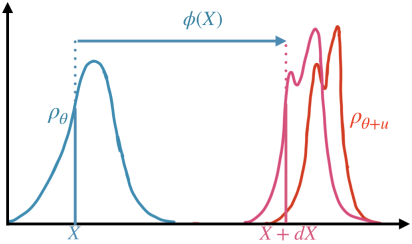
Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$



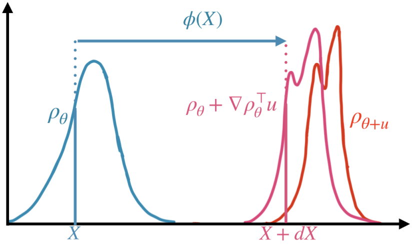
Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$



Wasserstein Natural Gradient: Step 2

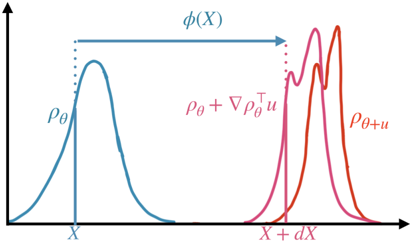
2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$



Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

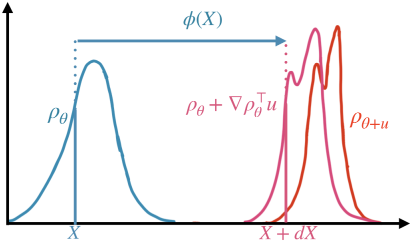
Elliptic equation: $\nabla \rho_\theta^\top u + div(\rho_\theta \phi) = 0$



Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

Elliptic equation: $\nabla\rho_\theta^\top u + div(\rho_\theta\phi) = 0$

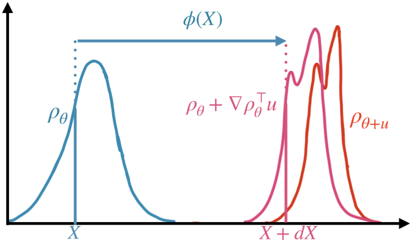


$$\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u}) \simeq \inf_\phi \int \|\phi(x)\|^2 \,\mathrm{d}\rho_\theta(x)$$

Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

Elliptic equation: $\nabla\rho_\theta^\top u + div(\rho_\theta\phi) = 0$



$$W_2^2(\rho_\theta, \rho_{\theta+u}) \simeq \int \|\phi(x)\|^2 \, \mathrm{d}\rho_\theta(x)$$

ϕ constrained to be 'almost' a gradient of a real valued function.

Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u}) \simeq \frac{1}{2}\int \|\phi(x)\|^2 \,\mathrm{d}\rho_\theta(x)$$
$$\nabla\rho_\theta^\top u + div(\rho_\theta\phi) = 0$$

Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u}) \simeq \frac{1}{2}\int \|\phi(x)\|^2 \,\mathrm{d}\rho_\theta(x)$$
$$\nabla\rho_\theta^\top u + div(\rho_\theta\phi) = 0$$

What does $\nabla\rho_\theta$ mean?

Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u}) \simeq \frac{1}{2}\int \|\phi(x)\|^2 \,\mathrm{d}\rho_\theta(x)$$
$$\nabla\rho_\theta^\top u + div(\rho_\theta\phi) = 0$$

What does $\nabla\rho_\theta$ mean?

$$\text{Implicit model} \Rightarrow \mathbb{E}_{\rho_\theta}[f(x)] = \int f(h_\theta(z))\,\mathrm{d}\nu(z)$$

Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u}) \simeq \frac{1}{2}\int \|\phi(x)\|^2 \, \mathrm{d}\rho_\theta(x)$$

$$\nabla\rho_\theta^\top u + div(\rho_\theta \phi) = 0$$

What does $\nabla\rho_\theta$ mean?

$$\text{Implicit model} \Rightarrow \mathbb{E}_{\rho_\theta}[f(x)] = \int f(h_\theta(z)) \, \mathrm{d}\nu(z)$$

Taylor expansion in u:

Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u}) \simeq \frac{1}{2}\int \|\phi(x)\|^2 \, \mathrm{d}\rho_\theta(x)$$

$$\nabla\rho_\theta^\top u + div(\rho_\theta \phi) = 0$$

What does $\nabla\rho_\theta$ mean?

$$\text{Implicit model} \Rightarrow \mathbb{E}_{\rho_\theta}[f(x)] = \int f(h_\theta(z)) \, \mathrm{d}\nu(z)$$

Taylor expansion in u:

$$\mathbb{E}_{\rho_{\theta+u}}[f(x)] - \mathbb{E}_{\rho_\theta}[f(x)] \simeq \int \nabla_x f(h_\theta(z)) \nabla h_\theta(z)^\top u \, \mathrm{d}\nu(z)$$

Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u}) \simeq \frac{1}{2}\int \|\phi(x)\|^2 \, \mathrm{d}\rho_\theta(x)$$

$$\nabla\rho_\theta^\top u + div(\rho_\theta \phi) = 0$$

What does $\nabla\rho_\theta$ mean?

$$\text{Implicit model} \Rightarrow \mathbb{E}_{\rho_\theta}[f(x)] = \int f(h_\theta(z)) \, \mathrm{d}\nu(z)$$

Taylor expansion in u:

$$\mathbb{E}_{\rho_{\theta+u}}[f(x)] - \mathbb{E}_{\rho_\theta}[f(x)] \simeq \underbrace{\int \nabla_x f(h_\theta(z)) \nabla h_\theta(z)^\top u \, \mathrm{d}\nu(z)}_{\nabla\rho_\theta(f)^\top u}$$

Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u}) \simeq \frac{1}{2}\int \|\phi(x)\|^2 \,\mathrm{d}\rho_\theta(x)$$

Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u}) \simeq \frac{1}{2}\int \|\phi(x)\|^2 \,\mathrm{d}\rho_\theta(x)$$

Variational expression for elliptic equations:

$$\nabla\rho_\theta^\top u + div(\rho_\theta \phi) = 0$$

$$\Updownarrow$$

$$\phi \in \arg \sup_{f \in C_c^\infty(\Omega)} \nabla\rho_\theta(f)^\top u - \frac{1}{2}\mathbb{E}_{\rho_\theta}[\|\nabla f(x)\|^2]$$

Wasserstein Natural Gradient: Step 2

2. Taylor expansion of $\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u})$

$$\frac{1}{2}W_2^2(\rho_\theta, \rho_{\theta+u}) \simeq \sup_{f \in C_c^\infty(\Omega)} \nabla\rho_\theta(f)^\top u - \frac{1}{2}\mathbb{E}_{\rho_\theta}[\|\nabla_x f(x)\|^2]$$

Variational expression for elliptic equations:

$$\nabla\rho_\theta^\top u + div(\rho_\theta \phi) = 0$$

$$\Updownarrow$$

$$\phi \in \arg \sup_{f \in C_c^\infty(\Omega)} \nabla\rho_\theta(f)^\top u - \frac{1}{2}\mathbb{E}_{\rho_\theta}[\|\nabla f(x)\|^2]$$

General recipe for natural gradients

1. Choose a distance/divergence d defined on the model ρ_θ:

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + d(\rho_{\theta_t+u}, \rho_{\theta_t})$$

2. Second order expansion of d at ρ_θ:

$$d(\rho_{\theta_t+u}, \rho_{\theta_t}) \simeq \frac{1}{2} u^\top G(\theta_t) u$$

3. Solve:

$$\min_u \nabla \mathcal{L}(\rho_{\theta_t})^\top u + \frac{1}{2} u^\top G(\theta_t) u$$

Wasserstein Natural Gradient: Step 3

3. Solve:

$$\min_{u} \nabla \mathcal{L}(\rho_\theta)^\top u + \sup_{f \in C_c^\infty(\Omega)} \nabla \rho_\theta(f)^\top u - \frac{1}{2}\mathbb{E}_{\rho_\theta}[\|\nabla_x f(x)\|^2]$$

Wasserstein Natural Gradient: Step 3

3. Solve:

$$\min_{u} \sup_{f \in C_c^\infty(\Omega)} \left(\nabla \mathcal{L}(\rho_\theta) + \nabla \rho_\theta(f)\right)^\top u - \frac{1}{2}\mathbb{E}_{\rho_\theta}[\|\nabla_x f(x)\|^2]$$

Wasserstein Natural Gradient: Step 3

3. Solve:

$$\min_{u} \sup_{f \in C_c^\infty(\Omega)} (\overbrace{\nabla \mathcal{L}(\rho_\theta) + \nabla \rho_\theta(f)}^{\mathcal{U}_\theta(f)})^\top u - \frac{1}{2}\mathbb{E}_{\rho_\theta}[\|\nabla_x f(x)\|^2]$$

Wasserstein Natural Gradient: Step 3

3. Solve:

$$\min_{u} \sup_{f \in C_c^\infty(\Omega)} \mathcal{U}_\theta(f)^\top u - \frac{1}{2}\mathbb{E}_{\rho_\theta}[\|\nabla_x f(x)\|^2]$$

Kernelized Wasserstein Natural Gradient

$$\min_{u} \sup_{f \in C_c^\infty(\Omega)} \mathcal{U}_\theta(f)^\top u - \frac{1}{2}\mathbb{E}_{\rho_\theta}[\|\nabla_x f(x)\|^2]$$

[3] [Mroueh et al., 2019]

Kernelized Wasserstein Natural Gradient

$$\min_{u} \sup_{f \in C_c^\infty(\Omega)} \mathcal{U}_\theta(f)^\top u - \frac{1}{2}\mathbb{E}_{\rho_\theta}[\|\nabla_x f(x)\|^2]$$

- Functional optimization: hard in general

[3] [Mroueh et al., 2019]

Kernelized Wasserstein Natural Gradient

$$\min_{u} \sup_{f \in \mathcal{H}} \mathcal{U}_\theta(f)^\top u - \frac{1}{2}\mathbb{E}_{\rho_\theta}[\|\nabla_x f(x)\|^2]$$

- Functional optimization: hard in general
- Replace $C_c^\infty(\Omega)$ by a nicer space: an RKHS $\mathcal{H}$[3].

[3] [Mroueh et al., 2019]

Kernelized Wasserstein Natural Gradient

$$\min_{u} \sup_{f \in \mathcal{H}} \mathcal{U}_\theta(f)^\top u - \frac{1}{2}\mathbb{E}_{\rho_\theta}[\|\nabla_x f(x)\|^2] + \frac{1}{2}(\epsilon \|u\|^2 - \lambda \|f\|_{\mathcal{H}}^2)$$

- Functional optimization: hard in general
- Replace $C_c^\infty(\Omega)$ by a nicer space: an RKHS $\mathcal{H}$[3].
- Regularization terms $\Rightarrow$ Ensure strong convexity/concavity

[3] [Mroueh et al., 2019]

Kernelized Wasserstein Natural Gradient

$$\sup_{f \in \mathcal{H}} \min_{u} \; \mathcal{U}_\theta(f)^\top u - \frac{1}{2}\mathbb{E}_{\rho_\theta}[\|\nabla_x f(x)\|^2] + \frac{1}{2}(\epsilon\|u\|^2 - \lambda\|f\|_{\mathcal{H}}^2)$$

- Functional optimization: hard in general
- Replace $C_c^\infty(\Omega)$ by an RKHS $\mathcal{H}$ with kernel k[4].
- Regularization terms $\Rightarrow$ Ensure strong convexity/concavity
- Minimax Theorem[5] $\Rightarrow$ Exchange order of min and sup.

[4] [Mroueh et al., 2019]
[5] [Ekeland and Téman, 1999]

Kernelized Wasserstein Natural Gradient

$$\sup_{f \in \mathcal{H}} \min_{u} \mathcal{U}_\theta(f)^\top u - \frac{1}{2}\mathbb{E}_{\rho_\theta}[\|\nabla_x f(x)\|^2] + \frac{1}{2}(\epsilon \|u\|^2 - \lambda \|f\|^2_{\mathcal{H}})$$

- Functional optimization: hard in general
- Replace $C_c^\infty(\Omega)$ by an RKHS $\mathcal{H}$ with kernel k[4].
- Regularization terms $\Rightarrow$ Ensure strong convexity/concavity
- Minimax Theorem[5] $\Rightarrow$ Exchange order of min and sup.
- Optimal descent direction u^* given by:

$$u^* = -\frac{1}{\epsilon}\mathcal{U}_\theta(f^*)$$

$$f^* = \arg\min_{f \in \mathcal{H}} \mathbb{E}_{\rho_\theta}[\|\nabla_x f(x)\|^2] + \frac{1}{\epsilon}\|\mathcal{U}_\theta(f)\|^2 + \lambda \|f\|^2_{\mathcal{H}}$$

---

[4] [Mroueh et al., 2019]

[5] [Ekeland and Téman, 1999]

KWNG: Sample based version

Have some i.i.d. samples $(Z_n)_{1 \leq n \leq N}$ from ν and $X_n = h_\theta(Z_n)$:

$$\hat{u}^* = -\frac{1}{\epsilon}\widehat{\mathcal{U}}_\theta(\hat{f}^*),$$

$$\hat{f}^* = \arg\min_{f \in \mathcal{H}} \frac{1}{N}\sum_{n=1}^{N} \|\nabla_x f(X_n)\|^2 + \frac{1}{\epsilon}\|\widehat{\mathcal{U}}_\theta(f)\|^2 + \lambda\|f\|_{\mathcal{H}}^2$$

KWNG: Sample based version

Have some i.i.d. samples $(Z_n)_{1 \leq n \leq N}$ from ν and $X_n = h_\theta(Z_n)$:

$$\hat{u}^* = -\frac{1}{\epsilon}\widehat{\mathcal{U}}_\theta(\hat{f}^*),$$

$$\hat{f}^* = \arg\min_{f \in \mathcal{H}} \frac{1}{N}\sum_{n=1}^{N} \|\nabla_x f(X_n)\|^2 + \frac{1}{\epsilon}\|\widehat{\mathcal{U}}_\theta(f)\|^2 + \lambda \|f\|^2_{\mathcal{H}}$$

- Representer theorem $\Rightarrow$ Optimal solution $\hat{f}^*$ of the form:

$$\hat{f}^* = \sum_{n=1}^{N}\sum_{i=1}^{d} \beta_{n,i}\partial_i k(X_n, .)$$

KWNG: Sample based version

Have some i.i.d. samples $(Z_n)_{1\leq n\leq N}$ from ν and $X_n = h_\theta(Z_n)$:

$$\hat{u}^* = -\frac{1}{\epsilon}\widehat{\mathcal{U}}_\theta(\hat{f}^*),$$

$$\hat{f}^* = \arg\min_{f\in\mathcal{H}} \frac{1}{N}\sum_{n=1}^{N} \|\nabla_x f(X_n)\|^2 + \frac{1}{\epsilon}\|\widehat{\mathcal{U}}_\theta(f)\|^2 + \lambda\|f\|_{\mathcal{H}}^2$$

- Representer theorem $\Rightarrow$ Optimal solution $\hat{f}^*$ of the form:

$$\hat{f}^* = \sum_{n=1}^{N}\sum_{i=1}^{d} \beta_{n,i}\partial_i k(X_n, .)$$

- $\beta_{n,i}$ obtained by solving a linear system of size Nd!!

KWNG: Sample based version

Have some i.i.d. samples $(Z_n)_{1 \leq n \leq N}$ from ν and $X_n = h_\theta(Z_n)$:

$$\hat{u}^* = -\frac{1}{\epsilon}\widehat{\mathcal{U}}_\theta(\hat{f}^*),$$

$$\hat{f}^* = \arg\min_{f \in \mathcal{H}} \frac{1}{N}\sum_{n=1}^{N} \|\nabla_x f(X_n)\|^2 + \frac{1}{\epsilon}\|\widehat{\mathcal{U}}_\theta(f)\|^2 + \lambda\|f\|_{\mathcal{H}}^2$$

- Nystrom projections [6] $\Rightarrow$ Reduce computational cost:

$$\hat{f}_M^* = \sum_{m=1}^{M} \alpha_m \partial_{i_m} k(\widetilde{X}_m, .)$$

[6] [Rudi et al., 2015, Sutherland et al., 2017]

KWNG: Sample based version

Have some i.i.d. samples $(Z_n)_{1\leq n\leq N}$ from ν and $X_n = h_\theta(Z_n)$:

$$\hat{u}^* = -\frac{1}{\epsilon}\widehat{\mathcal{U}}_\theta(\hat{f}^*),$$

$$\hat{f}^* = \arg\min_{f\in\mathcal{H}} \frac{1}{N}\sum_{n=1}^{N} \|\nabla_x f(X_n)\|^2 + \frac{1}{\epsilon}\|\widehat{\mathcal{U}}_\theta(f)\|^2 + \lambda\|f\|^2_{\mathcal{H}}$$

- Nystrom projections [6] $\Rightarrow$ Reduce computational cost:

$$\hat{f}^*_M = \sum_{m=1}^{M} \alpha_m \partial_{i_m} k(\widetilde{X}_m, .)$$

M sub-samples from $(X_i)_{1\leq i\leq N}$

[6] [Rudi et al., 2015, Sutherland et al., 2017]

KWNG: Sample based version

Have some i.i.d. samples $(Z_n)_{1\leq n\leq N}$ from ν and $X_n = h_\theta(Z_n)$:

$$\hat{u}^* = -\frac{1}{\epsilon}\widehat{\mathcal{U}_\theta}(\hat{f}^*),$$

$$\hat{f}^* = \arg\min_{f\in\mathcal{H}} \frac{1}{N}\sum_{n=1}^{N} \|\nabla_x f(X_n)\|^2 + \frac{1}{\epsilon}\|\widehat{\mathcal{U}_\theta}(f)\|^2 + \lambda\|f\|_{\mathcal{H}}^2$$

- Nystrom projections [6] $\Rightarrow$ Reduce computational cost:

$$\hat{f}_M^* = \sum_{m=1}^{M} \alpha_m \partial_{i_m} k(\widetilde{X}_m, .)$$

Randomly sampled from $\{1, ..., d\}$

M sub-samples from $(X_i)_{1\leq i\leq N}$

[6] [Rudi et al., 2015, Sutherland et al., 2017]

KWNG: Sample based version

After some further calculations:

$$\widehat{\nabla^W \mathcal{L}(\theta)} = \frac{1}{\epsilon}\left(I - T^\top (TT^\top + \lambda \epsilon K + \epsilon CC^\top)^\dagger T \right) \widehat{\nabla \mathcal{L}(\theta)}$$

KWNG: Sample based version

After some further calculations:

$$\widehat{\nabla^W \mathcal{L}(\theta)} = \frac{1}{\epsilon}\left(I - T^\top (TT^\top + \lambda\epsilon K + \epsilon C C^\top)^\dagger T\right)\widehat{\nabla \mathcal{L}(\theta)}$$

$$C_{m,(n,i)} = \frac{1}{\sqrt{N}} \partial_{i_m} \partial_{i+d} k(\widetilde{X}_m, X_n)$$

KWNG: Sample based version

After some further calculations:

$$K_{m,m'} = \partial_{i_m}\partial_{i_{m'}+d} k(\widetilde{X}_m, \widetilde{X}_{m'})$$

$$\widehat{\nabla^W \mathcal{L}(\theta)} = \frac{1}{\epsilon}\left(I - T^\top (TT^\top + \lambda\epsilon K + \epsilon C C^\top)^\dagger T\right) \widehat{\nabla \mathcal{L}(\theta)}$$

$$C_{m,(n,i)} = \frac{1}{\sqrt{N}} \partial_{i_m}\partial_{i+d} k(\widetilde{X}_m, X_n)$$

KWNG: Sample based version

After some further calculations:

$$K_{m,m'} = \partial_{i_m}\partial_{i_{m'}+d}k(\widetilde{X}_m, \widetilde{X}_{m'})$$

$$\widehat{\nabla^W \mathcal{L}(\theta)} = \frac{1}{\epsilon}\left(I - \boxed{T^\top}(TT^\top + \lambda\epsilon \boxed{K} + \epsilon \boxed{C} C^\top)^\dagger T\right)\widehat{\nabla \mathcal{L}(\theta)}$$

$$C_{m,(n,i)} = \frac{1}{\sqrt{N}}\partial_{i_m}\partial_{i+d}k(\widetilde{X}_m, X_n)$$

$T := \nabla\tau(\theta)$ with $\tau(\theta)_m = \frac{1}{N}\sum_{n=1}^{N}\partial_{i_m}k(\widetilde{X}_m, h_\theta(Z_n))$

Theory: Consistency and convergence rates

Theorem

Let δ be such that $0 \leq \delta \leq 1$. Under smoothness assumptions on the model characterized by some constant $c \geq 0$, for N large enough, $M \sim (dN^{\frac{2+c}{4+c}}\log(N))$, $\lambda \sim N^{\frac{1}{2b+1}}$ and $\epsilon \lesssim N^{-\frac{1}{4+c}}$, it holds with probability at least $1-\delta$ that:

$$\|\widehat{\nabla^W \mathcal{L}(\theta)} - \nabla^W \mathcal{L}(\theta)\|^2 = \mathcal{O}\Big(N^{-\frac{2}{4+c}}\Big).$$

KWNG: Ridgeless version

Additional structure when $\lambda = 0$:

$$\widehat{\nabla^{W}\mathcal{L}(\theta)} = \frac{1}{\epsilon}\left(I - T^{\top}(TT^{\top} + \epsilon CC^{\top})^{\dagger} T\right)\widehat{\nabla\mathcal{L}(\theta)}$$

KWNG: Ridgeless version

Additional structure when $\lambda = 0$:

$$\widehat{\nabla^W \mathcal{L}(\theta)} = \frac{1}{\epsilon} \left(I - T^\top (TT^\top + \epsilon CC^\top)^\dagger T \right) \widehat{\nabla \mathcal{L}(\theta)}$$

$$T_m = \frac{1}{N} \sum_{n=1}^{N} \nabla_\theta \partial_{i_m} k(Y_m, h_\theta(Z_n))$$

KWNG: Ridgeless version

Additional structure when $\lambda = 0$:

$$\widehat{\nabla^W \mathcal{L}(\theta)} = \frac{1}{\epsilon}\left(I - T^\top (TT^\top + \epsilon CC^\top)^\dagger T\right)\widehat{\nabla \mathcal{L}(\theta)}$$

$$T = CB, \qquad B_n = \nabla_\theta h_\theta(Z_n)$$

KWNG: Ridgeless version

Additional structure when $\lambda = 0$:

$$\widehat{\nabla^W \mathcal{L}(\theta)} = \frac{1}{\epsilon}\left(I - B^\top C^\top (CBB^\top C^\top + \epsilon CC^\top)^\dagger CB\right)\widehat{\nabla \mathcal{L}(\theta)}$$

$$T = CB, \qquad B_n = \nabla_\theta h_\theta(Z_n)$$

KWNG: Ridgeless version

Additional structure when $\lambda = 0$:

$$\widehat{\nabla^W \mathcal{L}(\theta)} = \frac{1}{\epsilon}\left(I - B^\top C^\top (CBB^\top C^\top + \epsilon CC^\top)^\dagger CB\right)\widehat{\nabla \mathcal{L}(\theta)}$$

$$T = CB, \qquad B_n = \nabla_\theta h_\theta(Z_n)$$

'Simplify' C:

$$\widetilde{T} = S^\dagger U^\top T, \qquad P = S^\dagger S$$

where $CC^\top = USU^\top$

KWNG: Ridgeless version

Additional structure when $\lambda = 0$:

$$\widehat{\nabla^W \mathcal{L}(\theta)} = \frac{1}{\epsilon}\left(I - \widetilde{T}^\top(\widetilde{T}\widetilde{T}^\top + \epsilon P)^\dagger \widetilde{T}\right)\widehat{\nabla \mathcal{L}(\theta)}$$

$$T = CB, \qquad B_n = \nabla_\theta h_\theta(Z_n)$$

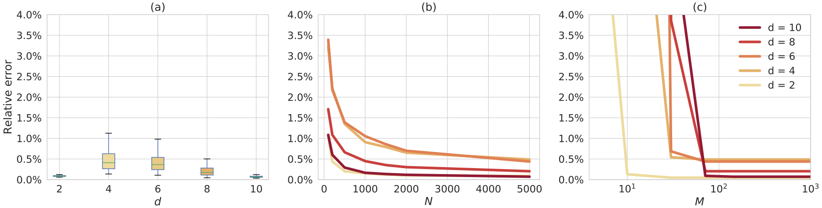
'Simplify' C:

$$\widetilde{T} = S^\dagger U^\top T, \qquad P = S^\dagger S$$

where $CC^\top = USU^\top$

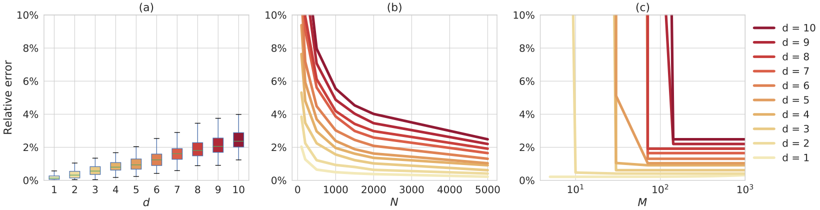
Experimental evaluation: Synthetic models

Hyper-spheres: $X = a + rZ, \qquad Z \sim \mathbb{S}_d$



Experimental evaluation: Synthetic models

$$\text{Gaussians: } X = a + rZ, \qquad Z \sim \mathcal{N}(0, I)$$

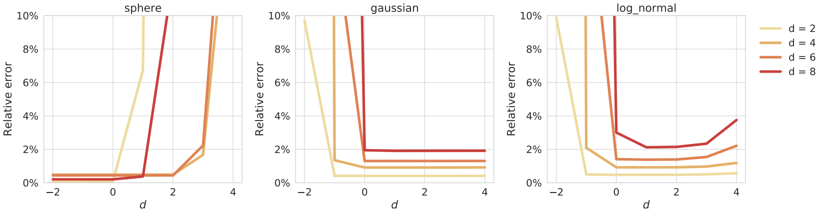


Experimental evaluation: Sensitivity to the choice of the kernel

- Gaussian kernel $k(x, y) = \exp(-\frac{\|x-y\|^2}{\sigma})$

Experimental evaluation: Sensitivity to the choice of the kernel

- Gaussian kernel $k(x, y) = \exp(-\frac{\|x-y\|^2}{\sigma})$

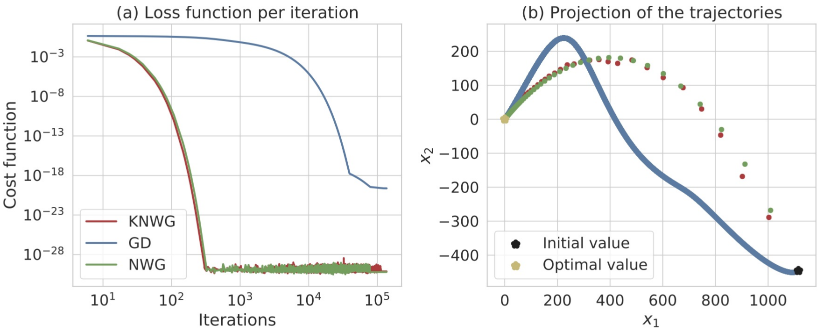


Experimental evaluation: Optimization trajectory

- Gaussian model for ρ_θ
- Loss functional $\mathcal{L}(\rho_\theta) = W_2^2(\rho_\theta, \rho_{\theta^*})$.

Experimental evaluation: Optimization trajectory

- Gaussian model for ρ_θ
- Loss functional $\mathcal{L}(\rho_\theta) = W_2^2(\rho_\theta, \rho_{\theta^*})$.



Experimental evaluation: Classification task

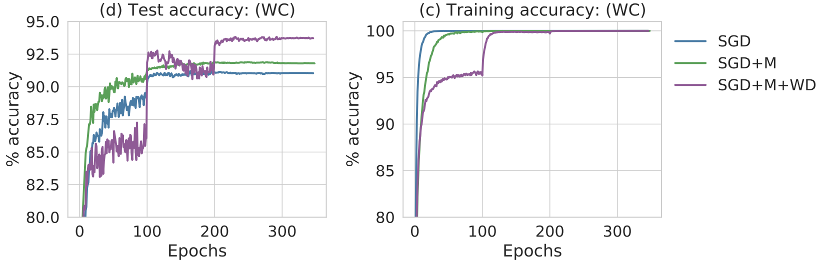
Well-conditioned problem:

$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(Z), Y) \, \mathrm{d}\nu(Z, Y)$$

Experimental evaluation: Classification task

Well-conditioned problem:

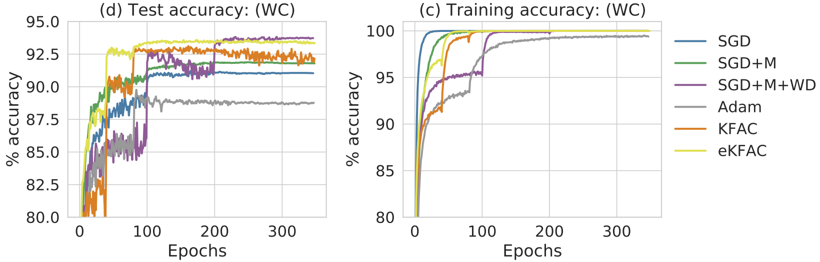
$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(Z), Y)\, \mathrm{d}\nu(Z, Y)$$



Experimental evaluation: Classification task

Well-conditioned problem:

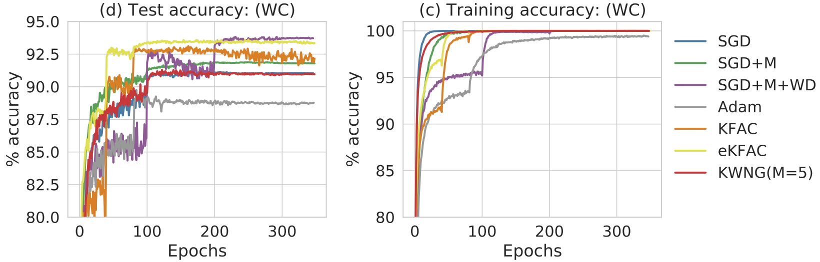
$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(Z), Y)\, \mathrm{d}\nu(Z, Y)$$



Experimental evaluation: Classification task

Well-conditioned problem:

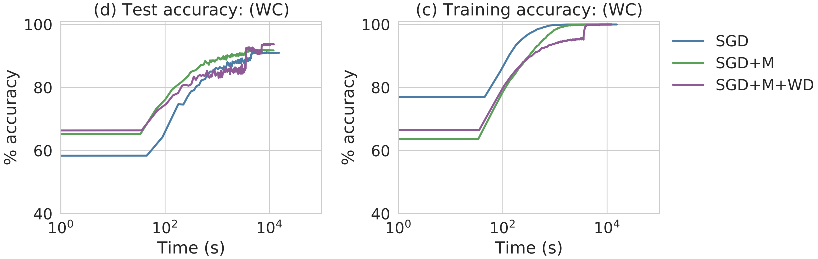
$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(Z), Y) \, \mathrm{d}\nu(Z, Y)$$



Experimental evaluation: Classification task

Well-conditioned problem:

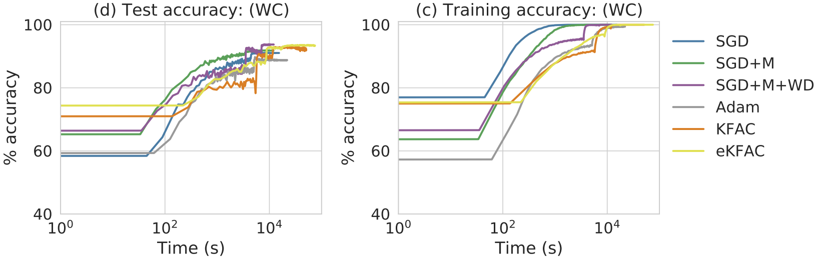
$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(Z), Y)\, \mathrm{d}\nu(Z, Y)$$



Experimental evaluation: Classification task

Well-conditioned problem:

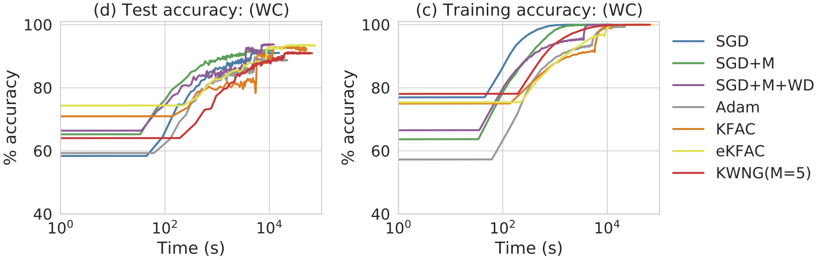
$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(Z), Y)\, \mathrm{d}\nu(Z, Y)$$



Experimental evaluation: Classification task

Well-conditioned problem:

$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(h_\theta(Z), Y) \, \mathrm{d}\nu(Z, Y)$$



Experimental evaluation: Classification task

Ill-conditioned problem:

$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(Uh_\theta(Z), Y)\,\mathrm{d}\nu(Z, Y)$$

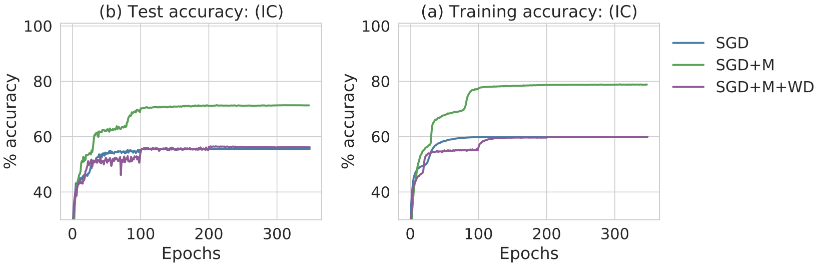
U is a diagonal matrix with $\kappa = 10^7$

Experimental evaluation: Classification task

Ill-conditioned problem:

$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(Uh_\theta(Z), Y)\, \mathrm{d}\nu(Z, Y)$$

U is a diagonal matrix with $\kappa = 10^7$

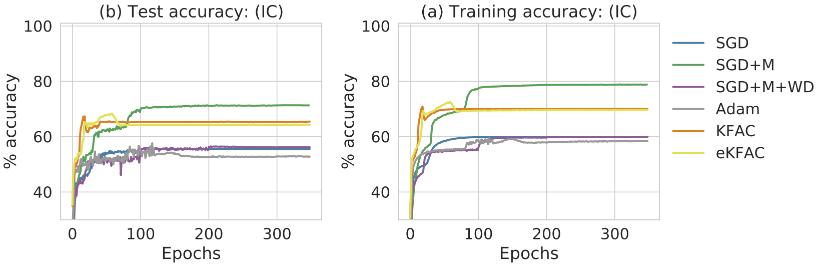


Experimental evaluation: Classification task

Ill-conditioned problem:

$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(Uh_\theta(Z), Y) \, \mathrm{d}\nu(Z, Y)$$

U is a diagonal matrix with $\kappa = 10^7$

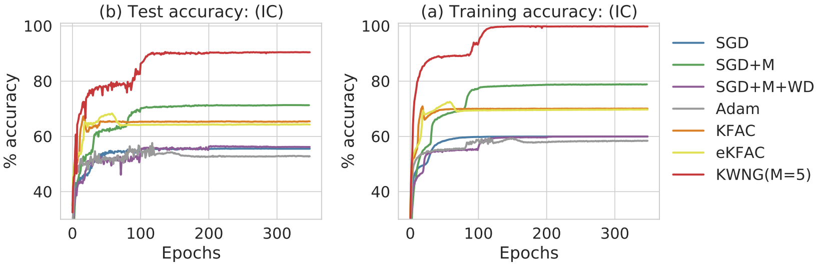


Experimental evaluation: Classification task

Ill-conditioned problem:

$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(Uh_\theta(Z), Y)\,\mathrm{d}\nu(Z, Y)$$

U is a diagonal matrix with $\kappa = 10^7$

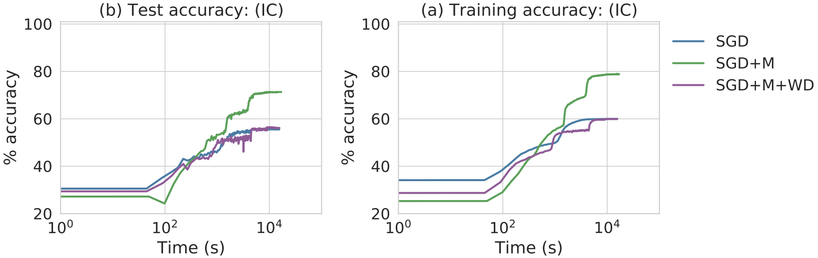


Experimental evaluation: Classification task

Ill-conditioned problem:

$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(Uh_\theta(Z), Y)\, \mathrm{d}\nu(Z, Y)$$

U is a diagonal matrix with $\kappa = 10^7$

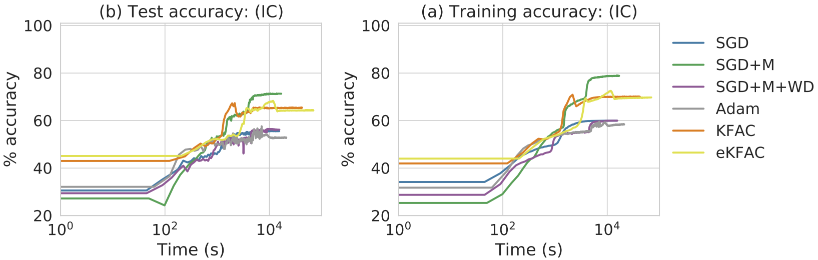


Experimental evaluation: Classification task

Ill-conditioned problem:

$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(Uh_\theta(Z), Y)\, \mathrm{d}\nu(Z, Y)$$

U is a diagonal matrix with $\kappa = 10^7$

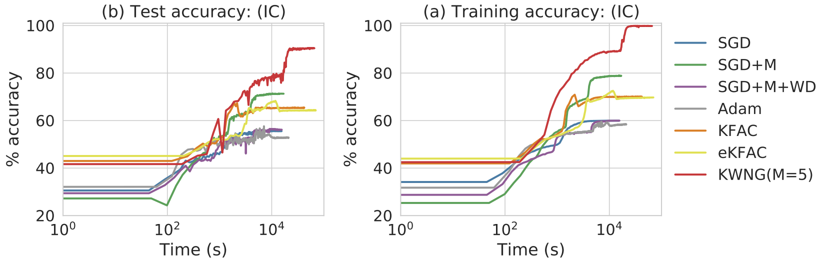


Experimental evaluation: Classification task

Ill-conditioned problem:

$$\min_{\theta} \mathcal{L}(\rho_\theta) := \int \ell(Uh_\theta(Z), Y)\,\mathrm{d}\nu(Z, Y)$$

U is a diagonal matrix with $\kappa = 10^7$



Conclusion

Summary of contributions

- Proposed to use the Wasserstein natural gradient for ill-conditioned problems
- A new algorithm to estimate the Wasserstein natural gradient
- Convergence rate: trade-off between computational complexity and statistical accuracy

Future work:

- Consistency result for the ridgeless version [7]
- Potential application in RL (implicit policy for RL [8])

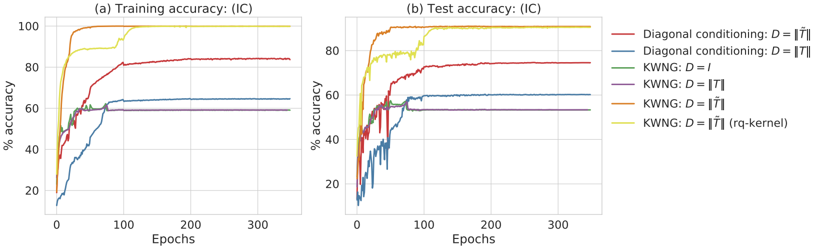
[7] [Liang et al., 2017]

[8] [Tang and Agrawal, 2019]

Thank you !

Ablation study

- Choice of the damping matrix $D(\theta)$
- Choice of the kernel (gaussian vs rational quadratic)



Benamou, J.-D. and Brenier, Y. (2000).
A computational fluid mechanics solution to the monge-kantorovich mass transfer problem.
Numerische Mathematik, 84(3):375–393.

Ekeland, I. and Téman, R. (1999).
Convex Analysis and Variational Problems.
Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

George, T., Laurent, C., Bouthillier, X., Ballas, N., and Vincent, P. (2018).
Fast Approximate Natural Gradient Descent in a Kronecker-factored Eigenbasis.
arXiv:1806.03884 [cs, stat].
arXiv: 1806.03884.

Grosse, R. and Martens, J. (2016).
A Kronecker-factored Approximate Fisher Matrix for Convolution Layers.
In *Proceedings of the 33rd International Conference on*