Data-Driven Geometry for Convex Optimisation

Hong Ye Tan

Joint work with: Carola-Bibiane Schönlieb, Subhadip Mukherjee, Junqi Tang, Andreas Hauptmann

Big Data Inverse Problems Workshop

23rd May 2024

UNIVERSITY OF CAMBRIDGE

Data-Driven Geometry for ~~Convex~~ Optimisation

Hong Ye Tan

Joint work with: Carola-Bibiane Schönlieb, Subhadip Mukherjee, Junqi Tang, Andreas Hauptmann

Big Data Inverse Problems Workshop

23rd May 2024

Motivation

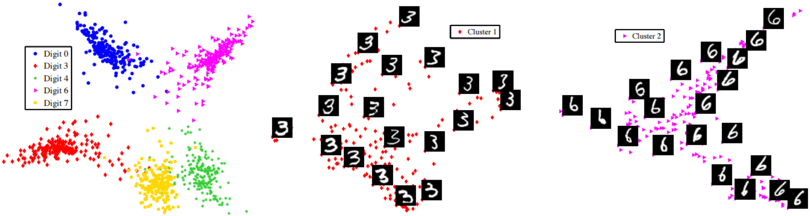
- Convex optimisation problems occur naturally in fields concerned with data analysis
- Reconstruct x from noisy measurement y

$$y = Ax + \varepsilon \rightsquigarrow \hat{x} = \arg\min_{x} \|Ax - y\|^2 + g(x)$$

- Do faster methods exist for specific classes of problems?
- **This talk:** Yes they do, we can learn them from data, and we can do so in a convergent manner.
    - **Learning to optimize:** optimization as a task

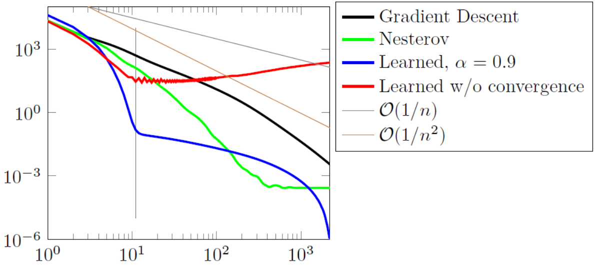
What is a “specific class”?

- No mathematical definition, only qualitative
- Problems are “similar”, e.g. forward operator, data type
- Examples: chest CT, natural image denoising
- Related: image manifold assumption



Background: learning to optimize

- Use neural network to parameterize update in terms of previous iterates
  - Ad-hoc convergence guarantees
- Parameterize as combination of proximal steps
  - Limited number of parameters
- **This work:** Convergent NN-based parameterization



[1]

[1] Banert et al., *Data-driven nonsmooth optimization*, SIAM Optimization, 2020

Background: Mirror Descent

Problem: Minimize convex function $f : \mathcal{X} = \mathbb{R}^n \to \mathbb{R}$

- Recall gradient descent with step-size η:

$$x_{k+1} = x_k - \eta \nabla f(x_k).$$

Ψ is $\mathcal{C}^1$ strongly convex, Ψ^* is the convex conjugate

Background: Mirror Descent

Problem: Minimize convex function $f : \mathcal{X} = \mathbb{R}^n \to \mathbb{R}$

- Recall gradient descent with step-size η:

$$x_{k+1} = x_k - \eta \nabla f(x_k).$$

- Issue: terms on RHS are not in the same space

$$x_{k+1} = \underbrace{x_k}_{\in \mathcal{X}} - \eta \underbrace{\nabla f(x_k)}_{\in \mathcal{X}^*}.$$

Ψ is $\mathcal{C}^1$ strongly convex, Ψ^* is the convex conjugate

Background: Mirror Descent

Problem: Minimize convex function $f : \mathcal{X} = \mathbb{R}^n \to \mathbb{R}$

- Recall gradient descent with step-size η:

$$x_{k+1} = x_k - \eta \nabla f(x_k).$$

- Issue: terms on RHS are not in the same space

$$x_{k+1} = \underbrace{x_k}_{\in \mathcal{X}} - \eta \underbrace{\nabla f(x_k)}_{\in \mathcal{X}^*}.$$

- Solution: have a (bijective) mirror map $\nabla \Psi : \mathcal{X} \to \mathcal{X}^*$, with inverse $(\nabla \Psi)^{-1} = \nabla \Psi^* : \mathcal{X}^* \to \mathcal{X}$

Ψ is $\mathcal{C}^1$ strongly convex, Ψ^* is the convex conjugate

Background: Mirror Descent

- This gives mirror descent (for strongly convex $\mathcal{C}^1$ Ψ):

$$x_{k+1} = (\nabla\Psi)^{-1}\left[\nabla\Psi(x_k) - \eta\nabla f(x_k)\right] \qquad \text{(MD)}$$

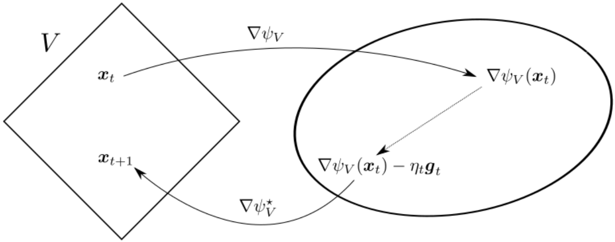


Figure: Schematic for MD[2]

[2] Image: F. Orabona. Online Mirror Descent II: Regret And Mirror Version.

Interpretations of MD

$$x_{k+1} = (\nabla\Psi)^{-1}\left[\nabla\Psi(x_k) - \eta\nabla f(x_k)\right] \tag{MD}$$

Interpretations of MD

$$x_{k+1} = (\nabla\Psi)^{-1}\left[\nabla\Psi(x_k) - \eta\nabla f(x_k)\right] \qquad \text{(MD)}$$

1. Proximal method with non-Euclidean divergence
   - GD: $x_{k+1} = \arg\min_x \left[\nabla f(x_k)^\top x + \frac{1}{2\eta}\|x - x_k\|_2^2\right]$
   - MD: $x_{k+1} = \arg\min_x \left[\nabla f(x_k)^\top x + \frac{1}{\eta}B_\Psi(x, x_k)\right]$

Interpretations of MD

$$x_{k+1} = (\nabla\Psi)^{-1}\left[\nabla\Psi(x_k) - \eta\nabla f(x_k)\right] \qquad \text{(MD)}$$

1. Proximal method with non-Euclidean divergence
   - GD: $x_{k+1} = \arg\min_x \left[\nabla f(x_k)^\top x + \frac{1}{2\eta}\|x - x_k\|_2^2\right]$
   - MD: $x_{k+1} = \arg\min_x \left[\nabla f(x_k)^\top x + \frac{1}{\eta}B_\Psi(x, x_k)\right]$
2. Weirdly-discretized Riemannian/preconditioned gradient flow

$$\dot{x} = -\left(\nabla^2\Psi(x)\right)^{-1}\nabla f(x) \qquad \text{(RGF)}$$

Interpretations of MD

$$x_{k+1} = (\nabla\Psi)^{-1}\left[\nabla\Psi(x_k) - \eta\nabla f(x_k)\right] \tag{MD}$$

1. Proximal method with non-Euclidean divergence
   - GD: $x_{k+1} = \arg\min_x \left[\nabla f(x_k)^\top x + \frac{1}{2\eta}\|x - x_k\|_2^2\right]$
   - MD: $x_{k+1} = \arg\min_x \left[\nabla f(x_k)^\top x + \frac{1}{\eta}B_\Psi(x, x_k)\right]$
2. Weirdly-discretized Riemannian/preconditioned gradient flow

$$\dot{x} = -\left(\nabla^2\Psi(x)\right)^{-1}\nabla f(x) \tag{RGF}$$

- Lower Lipschitz constant $\rightarrow$ larger step-size $\rightarrow$ faster convergence

Example: quadratic loss

- Optimizing $f(x) = 3x_1^2 + x_2^2$. "Optimal" $\Psi(x) = 9x_1^2 + x_2^2$.

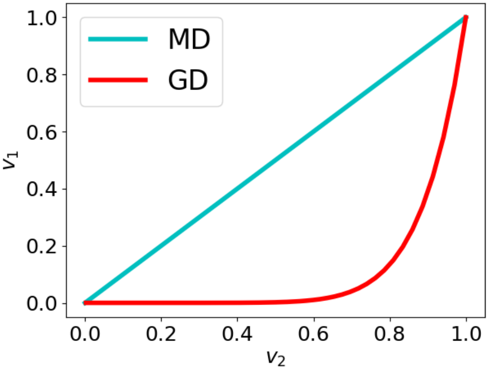


Figure: Optimization paths for MD and GD from $(1, 1)$. MD does not bend, allowing for larger step-size.

Classical convergence

Theorem (Informal)

Suppose $f : \mathcal{X} \to \mathbb{R}$ is convex, has L-Lipschitz gradient, and attains its minimizer in $\mathcal{X}$. Then for suitable step-size and mirror map, mirror descent has convergence rate

$$f(x_k) - f(x^*) = \mathcal{O}(1/k).$$

If additionally f is μ-strongly convex, mirror descent converges linearly:

$$f(x_k) - f(x^*) = \mathcal{O}\left(\left(1 + \frac{\mu}{L - \mu}\right)^{-k}\right).$$

Learning MD

MD: $x_{k+1} = (\nabla\Psi^*)\left[\nabla\Psi(x_k) - \eta\nabla f(x_k)\right].$

Learning MD

$$\text{MD: } x_{k+1} = (\nabla \Psi^*) \left[\nabla \Psi(x_k) - \eta \nabla f(x_k) \right].$$

$$\text{LMD: } \tilde{x}_{k+1} = (\nabla M_\vartheta^*) \left[\nabla M_\theta(\tilde{x}_k) - \eta \nabla f(\tilde{x}_k) \right].$$

- **Goal**: learn mirror maps $\nabla M_\theta \approx \nabla \Psi$, $\nabla M_\vartheta^* \approx \nabla \Psi^*$, where Ψ is the “optimal” mirror map for a given function class $\mathcal{F}$.

Learning MD

$$\text{MD: } x_{k+1} = (\nabla \Psi^*)\left[\nabla \Psi(x_k) - \eta \nabla f(x_k)\right].$$

$$\text{LMD: } \tilde{x}_{k+1} = (\nabla M_\vartheta^*)\left[\nabla M_\theta(\tilde{x}_k) - \eta \nabla f(\tilde{x}_k)\right].$$

- **Goal**: learn mirror maps $\nabla M_\theta \approx \nabla \Psi$, $\nabla M_\vartheta^* \approx \nabla \Psi^*$, where Ψ is the "optimal" mirror map for a given function class $\mathcal{F}$.

Classical	**Learned**
$\nabla \Psi^* = (\nabla \Psi)^{-1}$	$\nabla M_\vartheta^* \approx (\nabla M_\theta)^{-1}$
Ψ is strongly convex	M_θ, M_ϑ are strongly convex
Ψ is $\mathcal{C}^1$	M_θ, M_ϑ are $\mathcal{C}^1$

Convergence mechanism

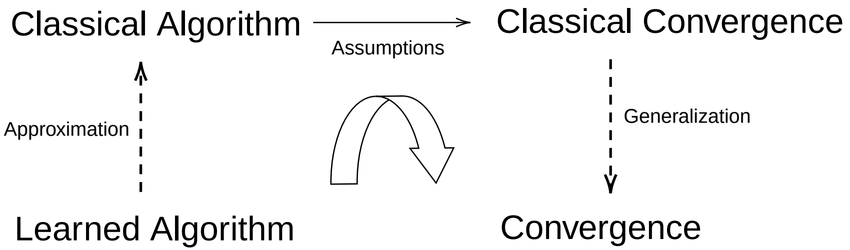
- How do we get convergence in the learned version?

Classical Algorithm $\xrightarrow[\text{Assumptions}]{}$ Classical Convergence

Learned Algorithm $\xrightarrow{?}$ Convergence?

Convergence mechanism

- How do we get convergence in the learned version?
- **A.** Modify the classical MD convergence results to the "approximate MD" case.



LMD Convergence guarantees

Theorem (Informal)

Let f be relatively L-smooth with respect to the mirror map Ψ. Suppose the approximation error

$$L\langle \nabla\Psi(x_i) - \nabla\Psi(\tilde{x}_i), x - \tilde{x}_i\rangle + \langle \nabla f(x_i), \tilde{x}_i - x_i\rangle \tag{1}$$

is uniformly bounded (above) by M. Approximate MD satisfies

$$\min_{1\leq i\leq k} f(\tilde{x}_i) - f(x) = \mathcal{O}(1/k) + M.$$

If f is also relatively μ-strongly convex with respect to Ψ,

$$\min_{1\leq i\leq k} f(\tilde{x}_i) - f(x) = \mathcal{O}\left(c^{-k}\right) + M.$$

Training objective

LMD goals (1) and (2) for a class of functions $\mathcal{F}$:

(1). Minimize objective functions f as quickly as possible;

(2). Enforce $\nabla M_\vartheta^* \approx (\nabla M_\theta)^{-1}$ by minimizing $\|\nabla M_\vartheta^* \circ \nabla M_\theta - I\|$.

$\Longrightarrow$ Training objective:

$$\tilde{x}_{k+1} = \nabla M_\vartheta^* \left(\nabla M_\theta(\tilde{x}_k) - t_k \nabla f(\tilde{x}_k)\right);$$

$$\mathcal{L}(\theta, \vartheta) = \sum_{f \in \mathcal{F}} \underbrace{\mathbb{E}\left[f(\tilde{x}_N)\right]}_{(1)} + \underbrace{\mathbb{E}_{\mathcal{X}}\left[\|\nabla M_\vartheta^* \circ \nabla M_\theta - I\|\right]}_{(2)}.$$

Example: Inpainting

- STL-10 dataset, 96×96 colour images
- Corrupted y using mask M with 20% missing pixels, 5% Gaussian noise
- Inpaint using TV regularization:

$$\min_x f(x; y) = \|M \circ (x - y)\|^2_{\mathcal{X}} + \lambda \|\nabla x\|_{1,\mathcal{X}}$$

- Function class[3] to learn LMD on:

$$\mathcal{F} = \{f(x; y) \mid \text{corrupted images } y\}$$

[3] This is split into training and testing subsets.

It’s fast

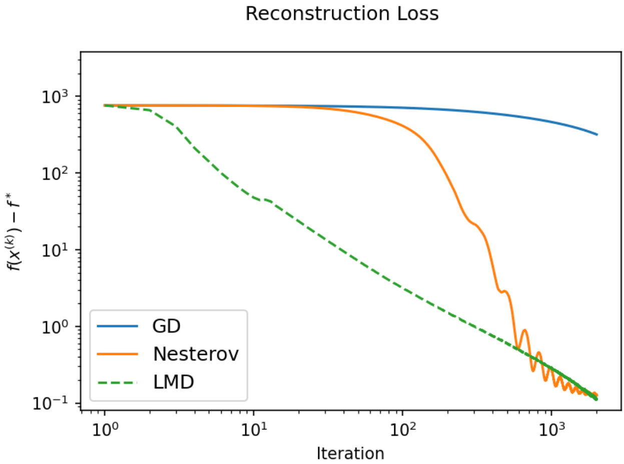


Figure: Much faster at small iterations

On unseen data (in test set).

Sanity check

TV-based reconstructions. Left to right: masked image, learned MD reconstruction, Adam based reconstruction.

What is it doing?

- Seems to “invert” the gradient at edges - sharpening?

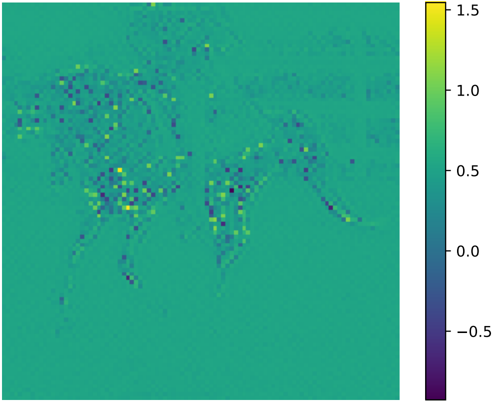


Figure: Pixel-wise $\nabla\Psi(y)/y$ (red channel)

It can be faster

- **Recent work:** It turns out we can accelerate LMD and also add stochasticity!
- Same pipeline: replace mirror maps in AMD with learned versions
- Convergence theory: similar to that of AMD
  - Convergence of accelerated LMD is to the minimum instead of minimum plus constant

Even faster

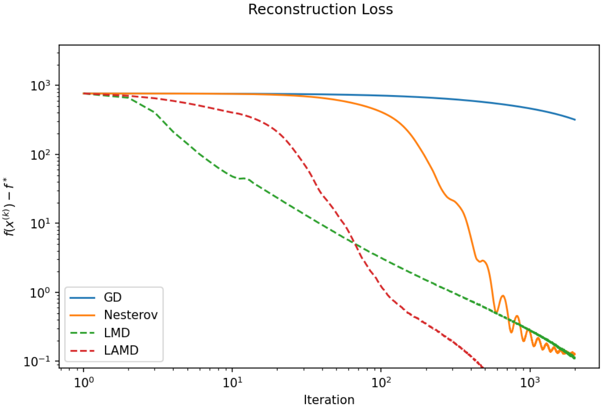
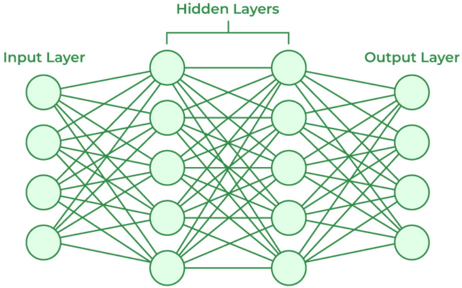


Figure: Reconstruction loss

Extension to non-convex NN training

- General idea: permuting intermediate features does not affect the final neural network (as a function) (*invariance*)
- Therefore, each individual element should be treated similarly to others in the same layer
- Allows for a layer-wise parameterization



Equivariance of L2O

Proposition

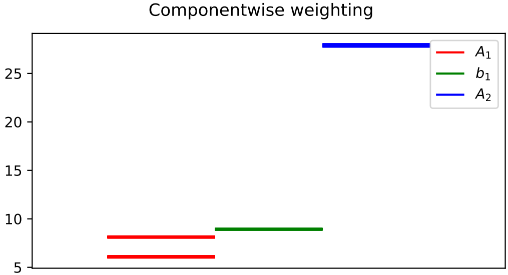
Let $(\mathcal{Z}, \langle \cdot, \cdot \rangle)$ be a Hilbert parameter space. Suppose that group G acts on $\mathcal{Z}$ linearly, such that

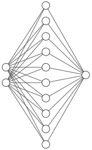
1. *The loss function $L : \mathcal{Z} \to \mathbb{R}$ is stable under G, that is, $L(g \cdot z) = L(z)$ for any $g \in G$ and $z \in \mathcal{Z}$;*
2. *The laws $p(z^{(0)})$ and $p(g \cdot z^{(0)})$ coincide for any $g \in G$.*

Then starting from a G-equivariant optimizer, a learned optimizer will continue to be G-equivariant.

Example: Weighted ℓ_2 potential

- Effectively give each element its own step-size (diagonal preconditioning).
- LMD Problem: Train a 1-hidden-layer neural network to classify 2D moons data (faster)





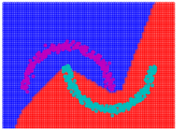


Figure: We observe that the LMD weights for the second layer matrix A_2 are almost constant. We see 2 bands for first matrix layer A_1 from the 2 input dimensions.

Initial experiments

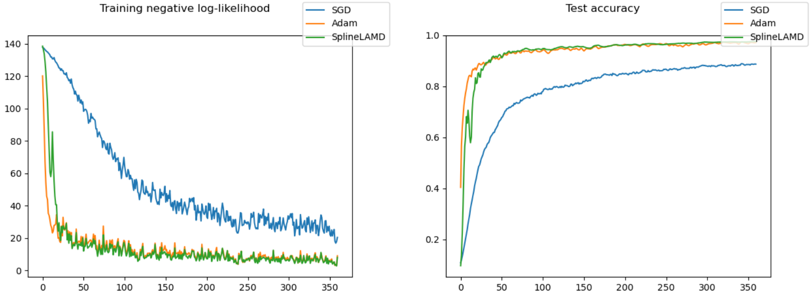


Figure: Comparison of training a four-hidden-layer neural network with SGD, Adam, and accelerated LMD for MNIST classification.

- LAMD is able to achieve very close performance to Adam (with different generalization performance!)

LMD: Summary

- MD: utilizing problem geometry $\rightarrow$ faster optimization
- LMD: data-driven geometry[4]
- Free equivariance for L2O![5]

Outlook

- Interpretability
- Optimal mirror maps?
- Characterising "smallness" of a class of functions

[4] HYT, Mukherjee, Tang, Schönlieb. *Data-driven mirror descent with input-convex neural networks*. SIMODS, 2023.

[5] HYT, Mukherjee, Tang, Schönlieb. *Boosting data-driven mirror descent with randomization, equivariance, and acceleration*. TMLR, 2024.

Definition of a derivative

Definition

A function $f : U \to \mathbb{R}$ is differentiable at $x \in U$ if there exists a linear map $A : \mathbb{R}^d \to \mathbb{R}$ such that for every h,

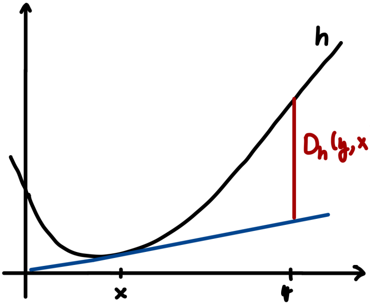
$$\lim_{t \to 0} \frac{f(x + th) - f(x) - tA(h)}{t} = 0.$$

We write $A = Df(x) \in B(\mathbb{R}^d, \mathbb{R})$.

Bregman divergence

$$B_h(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle \tag{2}$$

for convex distance generating function $h : \mathcal{X} \to \mathbb{R}$



Assumptions for MD

- Standard choices for $\Psi : \mathbb{R}^n \to \mathbb{R}$: strongly convex $\mathcal{C}^1$

Assumptions for MD

- Standard choices for $\Psi : \mathbb{R}^n \to \mathbb{R}$: strongly convex $\mathcal{C}^1$
- Convex conjugate $\Psi^*(p) = \sup_{x \in \mathbb{R}^d} \{\langle p, x \rangle + f(x)\}$
  - $\nabla \Psi^* = (\nabla \Psi)^{-1}$
- MD utilizes geometry of the problem
- Lower Lipschitz constant $\to$ larger allowed step-size $\to$ faster convergence

Example: simplex

Example

KL divergence on the simplex $\Delta = \{x \in \mathbb{R}^d : x \geq 0, \ \sum_i x_i = 1\}$

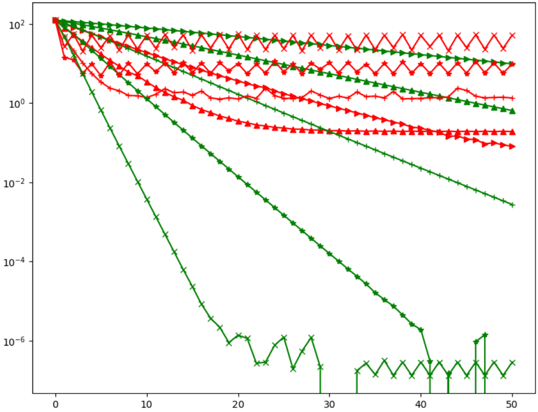
$$\min_{x \in \Delta} KL(x \| y) = \sum_{i=1}^{d} x_i \log\left(\frac{x_i}{y_i}\right)$$

Probabilistic distance: negative entropy

$$\Psi(x) = \sum_j x_j \log x_j \text{ if } x \in \Delta, \ +\infty \text{ otherwise}$$

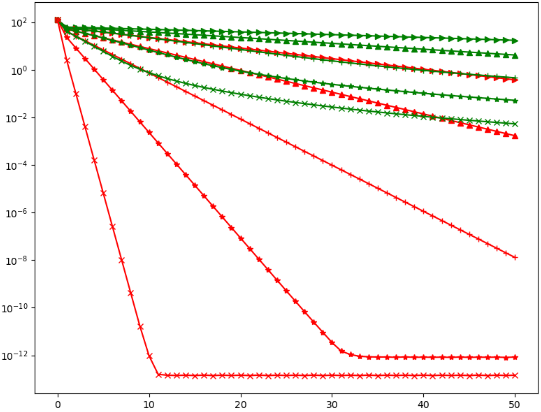
$$\nabla \Psi(x) = 1 + \log(x), \nabla \Psi^*(y) = \frac{\exp(y)}{\sum_j \exp(y_j)}$$

Example: KL on simplex



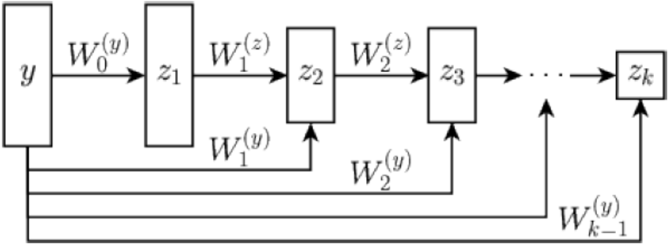
Green: MD with entropy function. Red: Projected subgradient descent

Example: least squares on simplex



Green: MD with entropy function. Red: Projected subgradient descent

ICNNs



Proposition

The function Ψ is convex in y if all $W_i^{(z)}$ are non-negative, and all functions g_i are convex and non-decreasing.

Convergence guarantees

Theorem (Formal)

Let f be relatively L-smooth and relatively μ-strongly-convex relative to the mirror map Ψ, with $L > 0$, $\mu \geq 0$. Consider the iterations

$$x_{k+1} = \arg\min_{x \in X} \left\{ \langle x, \nabla f(\tilde{x}_k) \rangle + L B_\Psi(x, \tilde{x}_k) \right\}, \quad \tilde{x}_{k+1} \approx x_{k+1}. \tag{3}$$

i.e. approximate MD with fixed step size $1/L$. Let $x \in \mathcal{X}$. Suppose

$$L \langle \nabla \Psi(x_i) - \nabla \Psi(\tilde{x}_i), x - \tilde{x}_i \rangle + \langle \nabla f(x_i), \tilde{x}_i - x_i \rangle \tag{4}$$

is uniformly bounded (above) by M. We have the following bound:

$$\min_{1 \leq i \leq k} f(\tilde{x}_i) - f(x) \leq \frac{\mu B_\Psi(x, \tilde{x}_0)}{(1 + \frac{\mu}{L-\mu})^k - 1} + M \leq \frac{L - \mu}{k} B_\Psi(x, \tilde{x}_0) + M. \tag{5}$$