SHANGHAI JIAO TONG UNIVERSITY

On Dynamics of Gradient Descent

Einstein, Rao and Nesterov

Yilin Zhang

September 10, 2024

Problem Formulation

One of the crucial problems to tackle in machine learning is how to optimize a function $f : \mathcal{M}_x \to \mathbb{R}$ with a metric ${}^{x}g$ defined on a n dimensional manifold $\mathcal{M}_x$.

$$\min_{x \in \mathcal{M}_x} f(x) \tag{1}$$

To clearly state the results, we make the following assumptions.

- f is m-strictly convex and ∇f is l-Lipschitz continuous.
- $h \in C^3$ is m_h-strictly convex.
- The distance between two points is measured by Bregman divergence B_h.

Assumption Explanation

Strongly Convex

A differentiable function f is called m-strongly convex if $\exists m > 0$ s.t. $\forall x, y$

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + m||y - x||^2.$$

Lipschitz Continuous

A function f is called l-lipschitz continuous if $\exists l > 0$, s.t. $\forall x, y$

$$||f(y) - f(x)|| \leq l||y - x||.$$

Note: f is m-strictly convex and ∇f is l-Lipschitz continuous. It means f is locally upper and lower bounded by a quadratic form.

Development of First Order Methods I

- Vanilla Gradient Descent(VGD)

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

- Heavy Ball Method(HVM)

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

- Nesterov Methods(Nes-sc)

$$\begin{aligned} y_{k+1} &= x_k - s\nabla f(x_k), \\ x_{k+1} &= y_{k+1} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}(y_{k+1} - y_k). \end{aligned}$$

A Geometrical View of Gradient Methods

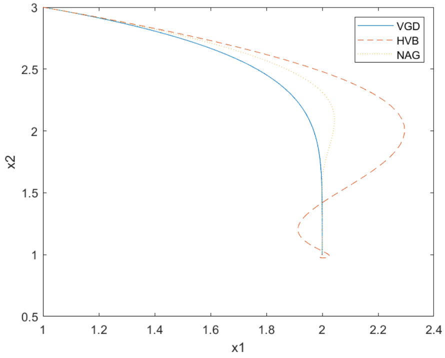

Figure: Path of Gradient Methods

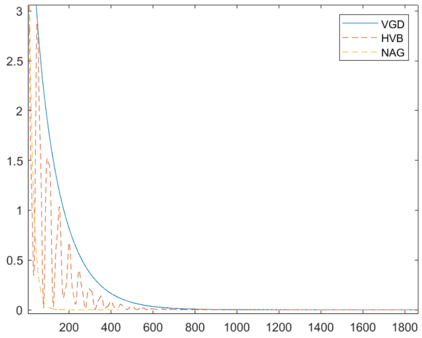

Figure: Convergence of Gradient Methods

Development of "First" Order Methods II

- Newton Nethod(NTM)

$$x_{k+1} = x_k - s\nabla^2 f(x_k)^{-1}\nabla f(x_k).$$

- Natural Gradient Descent(NGD)

$$x_{k+1} = x_k - s\mathcal{F}^{-1}\nabla f(x_k),$$

here $\mathcal{F}$ is Fisher matrix.

- Bregman Mirror Descent(BMD)

$$x_{k+1} = \arg\min_{x\in\mathcal{X}}\{\langle x, \nabla f(x_k)\rangle + sB_h(x, x_k)\},$$

here B_h is Bregman divergence.

Bregman divergence

Given a convex $h \in C^3$ as Bregman potential function, the Bregman divergence B_h is defined as

$$B_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle.$$

Table 1: Bregman divergences generated from some convex functions.

Domain	$\varphi(\mathbf{x})$	$d_\varphi(\mathbf{x}, \mathbf{y})$	**Divergence**
$\mathbb{R}$	x^2	$(x-y)^2$	Squared loss
$\mathbb{R}_+$	$x \log x$	$x \log(\frac{x}{y}) - (x-y)$	
$[0,1]$	$x \log x + (1-x)\log(1-x)$	$x \log(\frac{x}{y}) + (1-x)\log(\frac{1-x}{1-y})$	Logistic loss [3]
$\mathbb{R}_{++}$	$-\log x$	$\frac{x}{y} - \log(\frac{x}{y}) - 1$	Itakura-Saito distance
$\mathbb{R}$	e^x	$e^x - e^y - (x-y)e^y$	
$\mathbb{R}^d$	$\|\mathbf{x}\|^2$	$\|\mathbf{x}-\mathbf{y}\|^2$	Squared Euclidean distance
$\mathbb{R}^d$	$\mathbf{x}^T A \mathbf{x}$	$(\mathbf{x}-\mathbf{y})^T A (\mathbf{x}-\mathbf{y})$	Mahalanobis distance [4]
d-Simplex	$\sum_{j=1}^d x_j \log_2 x_j$	$\sum_{j=1}^d x_j \log_2(\frac{x_j}{y_j})$	KL-divergence
$\mathbb{R}_+^d$	$\sum_{j=1}^d x_j \log x_j$	$\sum_{j=1}^d x_j \log(\frac{x_j}{y_j}) - \sum_{j=1}^d (x_j - y_j)$	Generalized I-divergence

Geometrical Explanation

An explanation to bregman divergence is that it measures the distance between h and its linear approximation.

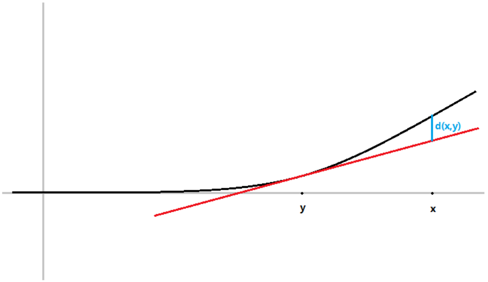

Figure: Geometrical Meaning of Bregman Divergence[1]

[1] Hanzhang Qin. *How to understand Bragman Divergence*. https://www.zhihu.com/question/22426561.

Consider that x_k is sampled from flow of $x(t)$ with step size δ. We take $x_k = x(\delta k) = x(t)$, where δ is the step size. Then we have the following Talyor series:

$$
\begin{aligned}
x_{k+1} &= x(k\delta + \delta) = x(t+\delta) = x(t) + \dot{x}(t)\delta + \frac{\ddot{x}(t)}{2}\delta^2 + O(\delta^3) \\
x_{k-1} &= x(k\delta - \delta) = x(t-\delta) = x(t) - \dot{x}(t)\delta + \frac{\ddot{x}(t)}{2}\delta^2 + O(\delta^3) \\
\nabla f(x_{k-1}) &= \nabla f(x(k\delta - \delta)) = \nabla f(x(t-\delta)) = \nabla f(x(t)) - \delta \nabla^2 f(x(t))\dot{x}(t) + O(\delta^2)
\end{aligned}
$$

Note: This is more like a reverse version of numerical solution.

From the form that $x_{k+1} = x_k - s\nabla f(x_k)$, we have that

$$\frac{d}{dt}(x(t)) = -\sqrt{s}\nabla f(x(t)),$$

where $\delta = \sqrt{s}$. Therefore, we have a vanilla gradient flow.

Vanilla Gradient Flow

$$\dot{x} = -\sqrt{s}\nabla f(x), \tag{2}$$

From the discretization $x_{k+1} = x_k - s\nabla f(x_k) + \alpha(x_k - x_{k-1})$ we have that

$$
\begin{aligned}
& x_{k+1} - x_k = \alpha(x_k - x_{k-1}) - s\nabla f(x_k) \\
\Leftrightarrow\ & \delta^2 \ddot{x}(t) + \frac{2(1-\alpha)}{1+\alpha}\delta \dot{x}(t) + \frac{2s}{1+\alpha}\nabla f(x(t)) + O(\delta^3) = 0
\end{aligned}
$$

If we take $\delta = \sqrt{\frac{2s}{1+\alpha}}$, the equation turns to be

$$
\begin{aligned}
& \ddot{x}(t) + \frac{2(1-\alpha)}{(1+\alpha)}\sqrt{\frac{1+\alpha}{2s}}\dot{x}(t) + \nabla f(x(t)) = 0 \\
\Rightarrow & \frac{d}{dt}\left(x + \frac{(1+\alpha)}{2(1-\alpha)}\sqrt{\frac{2s}{1+\alpha}}\dot{x}\right) = -\frac{(1+\alpha)}{2(1-\alpha)}\sqrt{\frac{2s}{1+\alpha}}\nabla f(x).
\end{aligned}
$$

High-resolution Flow of Nesterov

From the discretization that $y_{k+1} = x_k - s\nabla f(x_k), x_{k+1} = y_{k+1} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}(y_{k+1} - y_k)$. we have that

$$x_{k+1} = x_k - s\nabla f(x_k) + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}(x_k - s\nabla f(x_k) - x_{k-1} + s\nabla f(x_{k-1}))$$

$$\Leftrightarrow \delta^2 \ddot{x(t)} + 2\sqrt{\mu s}\delta \dot{x(t)} + (s - \sqrt{\mu}s^{\frac{3}{2}})\delta\nabla^2 f(X_t)\dot{x(t)} + s(1+\sqrt{\mu s})\nabla f(x(t)) = 0$$

If we take $\delta = \sqrt{s(1+\sqrt{\mu s})}$, then we get

$$\ddot{x(t)} + 2\sqrt{\frac{\mu}{1+\sqrt{\mu s}}}\dot{x(t)} + \sqrt{\frac{s}{1+\sqrt{\mu s}}}\nabla^2 f(x(t))\dot{x(t)} + \nabla f(x(t)) = 0$$

$$\Rightarrow \frac{d}{dt}\left(x + \frac{1}{2}\sqrt{\frac{1+\sqrt{\mu s}}{\mu}}\dot{x}\right) = -\frac{1}{2}\sqrt{\frac{1+\sqrt{\mu s}}{\mu}}\left(\nabla f(x) + \frac{s(1-\sqrt{\mu s})}{\sqrt{s(1+\sqrt{\mu s})}}\nabla^2 f(x)\dot{x}\right)$$

High-resolution Flow of HVB and Nes-sc

SHANGHAI JIAO TONG UNIVERSITY

Heavy Ball Flow

$$\frac{d}{dt}\left(x + \frac{(1+\alpha)}{2(1-\alpha)}\sqrt{\frac{2s}{1+\alpha}}\dot{x}\right) = -\frac{(1+\alpha)}{2(1-\alpha)}\sqrt{\frac{2s}{1+\alpha}}\nabla f(x) \tag{3}$$

Nesterov Flow

$$\frac{d}{dt}\left(x + \frac{1}{2}\sqrt{\frac{1+\sqrt{\mu s}}{\mu}}\dot{x}\right) = -\frac{1}{2}\sqrt{\frac{1+\sqrt{\mu s}}{\mu}}\left(\nabla f(x) + \frac{s(1-\sqrt{\mu s})}{\sqrt{s(1+\sqrt{\mu s})}}\nabla^2 f(x)\dot{x}\right) \tag{4}$$

An insight in information geometry shows that NTM, NGC, and BMD are equivalent.

Fisher Matrix and KL Divergence

Fisher information is the second derivative of KL divergence.

$$\mathcal{F}_\theta = \nabla^2_{\theta'} KL(\theta||\theta') \mid_{\theta'=\theta}$$

Note: The equation above shows that the Fisher matrix is actually some Hessian matrix of a function, the same as Newton's form $\nabla^2 f(x_k)$. Actually, there is an intrinsic quantity that reflects to Hessian form, which is the Riemannian metric.

Smooth Manifold

A smooth manifold is a topological space that is locally similar to euclidean space.

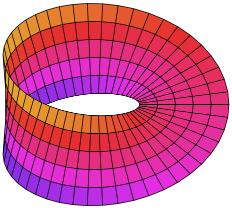

Figure: Möbius strip

Chart

A chart is a domain that maps part of the manifold. The cover of manifold with a collection of charts is called an atlas.

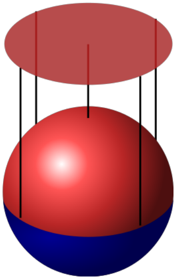

Figure: Sphere with Chart

- Chart allows us to map a surface to $\mathbb{R}^n$.
- A representation on a chart is a coordinate.

Directional Derivative

The directional derivative of a scalar function $f(x)$ along direction v is the function $D_v f(x)$ defined as

$$D_v f(x) = \lim_{h \to 0} \frac{f(x + hv) - f(x)}{h}.$$

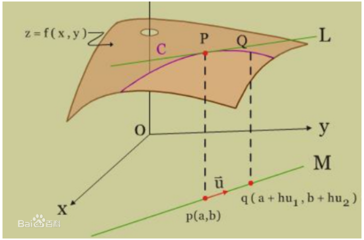

Figure: Directional Derivative

- Directional derivative is generalization of partial derivative.
- The direction depends on the direction chosen.

Tangent Plane

The tangent plane T_pM at point p is space spanned by tangent vectors.

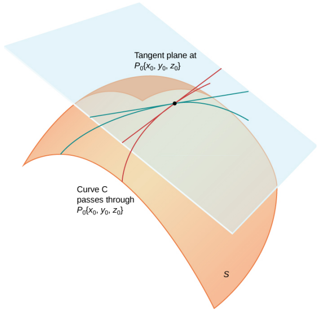

Figure: Tangent Plane

- Tangent plane is determined by surface parameterization.
- Directional derivative may not lay in tangent plane.

Covariant Derivative

A covariant derivative can be viewed as the orthogonal projection of the Euclidean directional derivative onto the manifold's tangent space. Let n be the surface normal.

$$\nabla_v f(x) = D_v f(x) - n.$$

Geodesic

A geodesic is defined as a curve $\gamma(t)$ such that parallel transport along the curve preserves the tangent vector to the curve on a manifold such that $\nabla_{\dot\gamma}\dot\gamma = 0$.

Note: Geodesic can also be derived by finding the infimin of $L = \int_a^b \sqrt{g_{\gamma(t)}(\dot\gamma, \dot\gamma)}$.

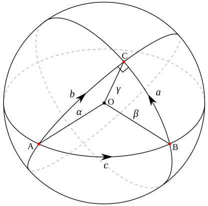

Figure: Geodesic on Sphere

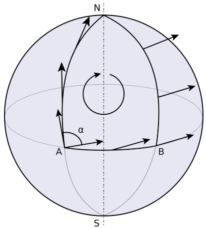

Figure: Parallel Transport along Geodesic

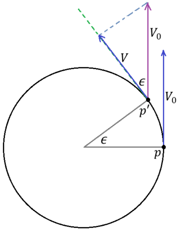

Figure: Parallel Transport on 1-D Sphere

Parallel Transport on Sphere

Riemannian Manifold

A Riemannian manifold (M, g) is a real, smooth manifold M equipped with a positive-definite inner product g_p on the tangent space T_pM at each point p.

Note: For 2-D surface embedded in $\mathbb{R}^3$, g is first fundamental form.

Levi-Civita Connection

The Levi-Civita connection is the unique affine connection on the tangent bundle of a manifold that preserves the Riemannian metric and is torsion-free. The Christoffel symbol of Levi-Civita connection is

$$\Gamma^{\mu}_{\lambda v} = \frac{1}{2} g^{\mu v} \left(\frac{\partial g_{\mu\lambda}}{\partial x^{v}} + \frac{\partial g_{v\mu}}{\partial x^{\lambda}} - \frac{\partial g_{\lambda v}}{\partial x^{v}} \right).$$

Here $\Gamma^{\mu}_{v\lambda}$ is the second Christoffel symbol.

Einstein notation

Take sum every repeated index in both upper and lower indices.

$$\sum_{i=1}^{3} c_i x^i \Rightarrow c_i x^i$$

Geodesic Equation

The geodesic equation defined on a Reimannian manifold with torsion free connection is

$$\frac{d^2 x^\mu}{d\tau^2} + \Gamma^\mu_{v\lambda} \frac{dx^v}{d\tau} \frac{dx^\lambda}{d\tau} = 0.$$

Next we are going to consider a special manifold consisting of distribution parameters.

Convex Dual

The dual Bregman divergence is defined as B_{h^*} where h^* is the convex dual defined as

$$h^*(\theta) = \sup_{x \in \mathcal{M}} \{\langle x, \theta \rangle - h(x)\}.$$

Fundamental Theorem of Information Geometry

If a torsion-free affine connection ∇ has constant curvature κ then its conjugate torsion-free connection ∇^* has necessarily the same constant curvature κ.

Dual Flat Structure of Bregman Statistical Manifold

A dually flat manifold $(\mathcal{M}_x, {}^h g, \nabla, \nabla^*)$ generated by a Bregman divergence B_h satisfies that ${}^h g = \nabla^2 h(x)$, $\mathcal{M}_x$ is both ∇-flat and ∇^*-flat. Here ∇^* is generated by ${}^h g^*$.

Note: An affine connection is said to be flat if its curvature tensor is 0.

NGD and BMD

Bregman mirror descent on the Hessian manifold $(M, g = \nabla^2 h(x))$ is equivalent to natural gradient descent on the dual Hessian manifold $(M, g = \nabla^2 h(\eta))$, where g is Bregman generator, $= \nabla h(x)$ and $x = \nabla h(\eta)$.

Dual Flat Manifold Geodesic

In a DFM, we have two global affine coordinate systems $x()$ and $()$ related by the Legendre-Fenchel transformation of a pair of potential functions h and h^*. That is, $(M,h)(M,h)$, and the dual atlases are $A = (M,x)$ and $A = (M,\eta)$. In a dually flat manifold, any pair of points P and Q can either be linked using the ∇-geodesic (that is x-straight) or the ∇-geodesic (that is -straight).

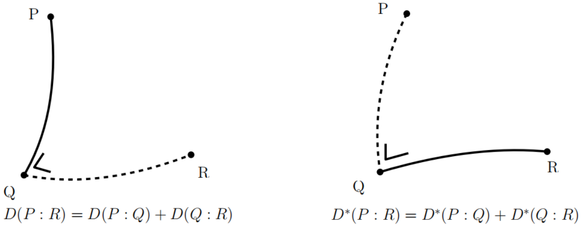

Figure: Dual Geodesic Connection

Consider a manifold $\mathcal{M} = \mathbb{R} \times \mathbb{R}^3$ such that $X = (t, x^1, x^2, x^3)$. Here t is real time, x^i is coordinate to describe position of particle. Then the equation describes the curvature of $\mathcal{M}$, energy and momentum is as below.

Einstein's Field Equation

R_{uv} is Ricci tensor, T_{uv} is the energy–momentum tensor, g_{uv} is Riemannian curvature.

$$R_{uv} - \frac{1}{2} R g_{uv} + \Lambda g_{uv} = \kappa T_{uv}$$

Note: The Newton limit has parameter: $\kappa = \frac{8\pi G}{c^4}$.

World Line

In general relativity, the world line of a particle free from all external, non-gravitational force is a particular type of geodesic in curved spacetime. In other words, a freely moving or falling particle always moves along a geodesic.

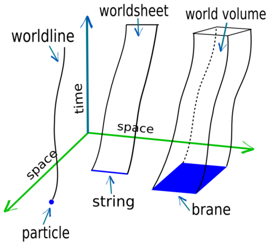

Figure: World Line Example

Newton's Law

Under the following three assumptions, the geodesic equation of a particle is reduced to Newton's law.

- Slow motion $|\frac{dX^i}{d\tau}| \ll |\frac{dX^0}{d\tau}|$.
- Static field $\frac{dg}{dX^0} = 0$.
- Weak field $g = \eta + \phi$. Here η is Minkowski metric, ϕ is a tiny disturbance.

Now consider the geodesic equation in $\mathcal{M}$ as:

$$\frac{d^2X^\mu}{d\tau^2} + \Gamma^\mu_{\gamma\lambda}\frac{dX^\gamma}{d\tau}\frac{dX^\lambda}{d\tau} = 0 \quad (\mu = 0, 1, 2, 3).$$

We consider "slow" motion of particle under "static" field. By "slow" we have $|\frac{dX^i}{d\tau}| \ll \frac{dX^0}{d\tau}$ and by "static" we have $\frac{\partial g}{\partial X^0} = 0$ This leads to

$$\frac{d^2X^\mu}{d\tau^2} + \Gamma^\mu_{00}(\frac{dX^0}{d\tau})^2 = 0$$

$$\frac{d^2X^\mu}{d\tau^2} - \frac{1}{2}g^{\mu v}(\frac{\partial g_{00}}{\partial X^v})(\frac{dX^0}{d\tau})^2 = 0 \quad (\mu = 0, 1, 2, 3).$$

When $\mu = 0$, since $\frac{\partial g_{00}}{\partial X^0} = 0$, we get $\frac{d^2X^0}{d\tau^2} = 0$. Solve it gives $t = c_1\tau + c_2$. With out loss of generality, we select $c_1 = -1, c_2 = 0$ and the $t = \tau$.

Now we consider $\mu = 1, 2, 3$. With $t = \tau$ we have

$$\frac{d^2 X^\mu}{dt^2} + \sum_{v=1}^{3} \frac{1}{2} g^{\mu v} \left(\frac{\partial g_{00}}{\partial X^v}\right) = 0 \quad (\mu = 1, 2, 3).$$

Apply approximation $g = \eta + \phi$ and omit $O(h^2)$ we get

$$\frac{d^2 X^\mu}{dt^2} + \frac{1}{2} \frac{\partial h_{00}}{\partial X^\mu} = 0 \quad (\mu = 1, 2, 3).$$

Then by taking $h = -2\Phi + c$, we get the Newton's law of a particle in 3-D space moving in potential field Φ as

$$\ddot{X} = \nabla\Phi$$

Now we repeat the above procedure for DFM. Take $\mathcal{M} = \mathbf{R} \times \mathcal{M}_x$ where $X = (t, x_1, ..., x_n)$. Consider the geodesic equation in $\mathcal{M}$ as:

$$\frac{d^2 X^\mu}{d\tau^2} + \Gamma^\mu_{\gamma\lambda} \frac{dX^\gamma}{d\tau} \frac{dX^\lambda}{d\tau} = 0 \quad (\mu = 0, 1, ..., n).$$

We consider "slow" motion of particle under "static" field. By "slow" we have $|\frac{dX^i}{d\tau}| \ll \frac{dX^0}{d\tau}$ and by "static" we have $\frac{\partial g}{\partial X^0} = 0$ This leads to

$$\frac{d^2 X^\mu}{d\tau^2} + \Gamma^\mu_{00} (\frac{dX^0}{d\tau})^2 = 0$$
$$\frac{d^2 X^\mu}{d\tau^2} - \frac{1}{2} g^{\mu v} (\frac{\partial g_{00}}{\partial X^v})(\frac{dX^0}{d\tau})^2 = 0 \quad (\mu = 0, 1, 2, ..., n).$$

When $\mu = 0$, since $\frac{\partial g_{00}}{\partial X^0} = 0$, we get $\frac{d^2 X^0}{d\tau^2} = 0$. Solve it gives $t = c_1 \tau + c_2$. With out loss of generality, we select $c_1 = 1, c_2 = 0$ and the $t = \tau$.

Now we consider $\mu = 1, ..., n$. With $t = \tau$ we have

$$\frac{d^2 X^\mu}{dt^2} - \sum_{v=1}^{n} \frac{1}{2} g^{\mu v} (\frac{\partial g_{00}}{\partial X^v}) = 0 \quad (\mu = 1, 2, ..., n).$$

Pile n equations we get

$$\frac{d^2}{dt^2} \begin{bmatrix} X^1 \\ \vdots \\ X^n \end{bmatrix} = \frac{1}{2} \begin{bmatrix} g^{11} & \cdots & g^{1n} \\ \vdots & \ddots & \vdots \\ g^{n1} & \cdots & g^{nn} \end{bmatrix} \begin{bmatrix} \frac{\partial g_{00}}{\partial X^1} \\ \vdots \\ \frac{\partial g_{00}}{\partial X^n} \end{bmatrix}$$

$$\Leftrightarrow \ddot{x} = \frac{1}{2} \nabla^2 h(x)^{-1} \nabla g_{00}$$

By taking $g_{00} = -2\Phi(x) + C$ we get that $\ddot{x} = -\nabla^2 h(x)^{-1} \nabla \Phi(x)$

Consider a damping system on the geodesic of DFM with $\Phi(x) = \gamma f(x) + \frac{d}{dt}(\alpha h(x) + \beta f(x))$. Then we have

$$\ddot{x} = -\nabla^2 h(x)^{-1}(\gamma \nabla f(x) + \nabla \frac{d}{dt}(\alpha h(x) + \delta f(x)))$$

$$\ddot{x} + \alpha \dot{x} = -\nabla^2 h(x)^{-1}(\gamma \nabla f(x) + \beta \nabla^2 f(x) \dot{x})$$

$$\nabla^2 h(x)(\ddot{x} + \alpha \dot{x}) + \beta \nabla^2 f(x) \dot{x} + \gamma \nabla f(x) = 0.$$

This is the unified form of all gradient based equations.

Theorem

Given a dually flat Bregman manifold $(\mathcal{M}_x, {}^h g, \nabla, \nabla^*)$ generated by Bregman divergence $B_h(\cdot, \cdot)$, the first order gradient methods are derived from the geodesic on the statistical manifold $(\mathcal{M}, g, \nabla)$ where $\mathcal{M} = \mathbb{R} \times \mathcal{M}_x$, ∇ is Levi-Civita connection and

$$g = \begin{bmatrix} -2f(x) + C & 0 \\ 0 & \nabla^2 h(x) \end{bmatrix}$$

Parameters of Gradient Flows

Method	α	β	δ	$h(\cdot)$
VGD	$\sqrt{\frac{2}{s}}$	0	$\sqrt{2s}$	$\frac{\lVert\cdot\rVert^2}{2}$
HVB	$\frac{2(1-\alpha)}{(1+\alpha)}\sqrt{\frac{1+\alpha}{2s}}$	0	$\sqrt{\frac{2s}{1+\alpha}}$	$\frac{\lVert\cdot\rVert^2}{2}$
Nes-sc	$2\sqrt{\frac{\mu}{1+\sqrt{\mu s}}}$	$\sqrt{\frac{s}{1+\sqrt{\mu s}}}(1-\sqrt{\mu s})$	$\sqrt{s(1+\sqrt{\mu s})}$	$\frac{\lVert\cdot\rVert^2}{2}$
NGD	$\sqrt{\frac{2}{s}}$	0	$\sqrt{2s}$	$h(\cdot)$
BMD	$\sqrt{\frac{2}{s}}$	0	$\sqrt{2s}$	$h^*(\cdot)$

What's Next

- Sensitivity analysis of parameters.
- A better explanation to the acceleration.

Q&A

Any question or suggestion is warmly welcomed.