# Information geometry of mirror descent

## Geometric Science of Information

Anthea Monod

Department of Statistical Science
Duke University Information Initiative at Duke

G. Raskutti (UW Madison) and S. Mukherjee (Duke)

29 Oct 2015

# Optimization of large-scale problems

Optimization of a function $f(\theta)$ where $\theta \in \mathbb{R}^p$.

$O(\sqrt{p})$ - convergence rate of standard subgradient descent. A problem in modern optimization, e.g. machine learning.

Mirror descent [A Nemirovski, 1979. A Beck & M Teboulle, 2003]:
$O(\log p)$ - convergence rate of mirror descent. Widely used tool in optimization and machine learning.

# Differential geometry in statistics

(1) Cramér-Rao lower bound (Rao 1945) - Lower bound on the variance of an estimator is a function of curvature. Sometimes called Cramér-Rao-Fréchet-Darmois lower bound.

(2) Invariant (non-informative) priors (Jeffreys 1946) - An uniformative prior distribution for a parameter space is based on a differential form.

(3) Information geometry (Amari 1985) - Differential geometry of probability distributions.

# Stochastic gradient descent

Given a convex differentiable cost function, $f : \Theta \to \mathbb{R}$.
Generate a sequence of parameters $\{\theta_t\}_{t=1}^{\infty}$ which incur a loss $f(\theta_t)$ that minimize *regret* at a time $T$, $\sum_{t=1}^{T} f(\theta_t)$.

One solution

$$\theta_{t+1} = \theta_t - \alpha_t \nabla f(\theta_t),$$

where $(\alpha_t)_{t=0}^{\infty}$ denotes a sequence of step-sizes.

# Natural gradient

For certain cost functions (log-likelihoods of exponential family models) the set of parameters $\Theta$ are supported on a $p$-dimensional Riemannian manifold, $(\mathcal{M}, \mathcal{H})$.
Typically the metric tensor $\mathcal{H} = (h_{jk})$ is determined by the Fisher information matrix

$$(\mathcal{I}(\theta))_{ij} = \mathbb{E}_{\text{Data}} \left[ \left( \frac{\partial}{\partial \theta_i} f(x;\theta) \right) \left( \frac{\partial}{\partial \theta_j} f(x;\theta) \right) \Big|_{\theta} \right], \quad i, j = 1, \dots, p.$$

# Natural gradient

Given a cost function $f$ on the Riemannian manifold $f : \mathcal{M} \to \mathbb{R}$, the *natural* gradient descent step is:

$$\theta_{t+1} = \theta_t - \alpha_t \mathcal{H}^{-1}(\theta_t) \nabla f(\theta_t),$$

where $\mathcal{H}^{-1}$ is the inverse of the Riemannian metric.

The natural gradient algorithm steps in the direction of steepest descent along the Riemannian manifold $(\mathcal{M}, \mathcal{H})$. It requires a matrix inversion.

# Mirror descent

Gradient descent can be written

$$\theta_{t+1} = \arg\min_{\theta \in \Theta} \left\{ \langle \theta, \nabla f(\theta_t) \rangle + \frac{1}{2\alpha_t} \|\theta - \theta_t\|_2^2 \right\}.$$

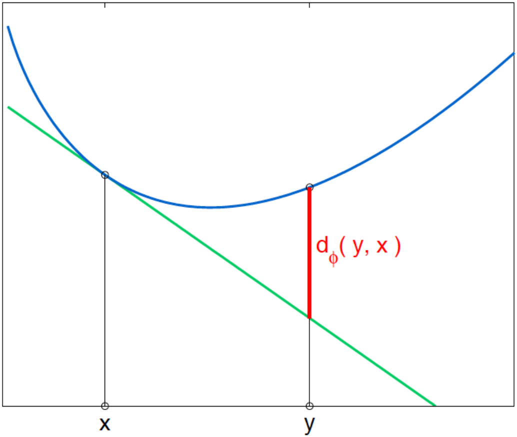
For a (strictly) convex proximity function $\Psi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$ mirror descent is

$$\theta_{t+1} = \arg\min_{\theta \in \Theta} \left\{ \langle \theta, \nabla f(\theta_t) \rangle + \frac{1}{\alpha_t} \Psi(\theta, \theta_t) \right\}.$$

# Bregman divergence

Let $G : \Theta \to \mathbb{R}$ be a strictly convex twice-differentiable function the Bregman divergence is

$$B_G(\theta, \theta') = G(\theta) - G(\theta') - \langle \nabla G(\theta'), \theta - \theta' \rangle.$$



# Bregman divergences for exponential family

| Family | $G(\theta)$ | $B_G(\theta, \theta')$ |
|---|---|---|
| $\mathcal{N}(\theta, I_{p\times p})$ | $\frac{1}{2}\|\theta\|_2^2$ | $\frac{1}{2}\|\theta - \theta'\|_2^2$ |
| $\mathrm{Poi}(e^\theta)$ | $\exp(\theta)$ | $\exp(\theta/\theta') - \langle \exp(\theta'), \theta - \theta' \rangle$ |
| $\mathrm{Be}\left(\frac{1}{1+e^{-\theta}}\right)$ | $\log(1+\exp(\theta))$ | $\log\left(\frac{1+e^\theta}{1+e^{\theta'}}\right) - \left\langle \frac{e^{\theta'}}{1+e^{\theta'}}, \theta - \theta' \right\rangle$ |

# Mirror descent

Mirror descent using the Bregman divergence as the proximity function

$$\theta_{t+1} = \arg\min_{\theta} \left\{ \langle \theta, \nabla f(\theta_t) \rangle + \frac{1}{\alpha_t} B_G(\theta, \theta_t) \right\}.$$

# Convex duals

The convex conjugate function for a function $G$ is defined to be:

$$H(\mu) := \sup_{\theta \in \Theta} \{\langle \theta, \mu \rangle - G(\theta)\} .$$

Let $\mu = g(\theta) \in \Phi$ be the extremal point of the dual. The dual Bregnman divergence $B_H : \Phi \times \Phi \to \mathbb{R}^+$ is

$$B_H(\mu, \mu') = H(\mu) - H(\mu') - \langle \nabla H(\mu'), \mu - \mu' \rangle.$$

# Dual Bregman divergences for exponential family

| $G(\theta)$ | $H(\mu)$ | $B_H(\mu, \mu')$ |
|---|---|---|
| $\frac{1}{2}\|\theta\|_2^2$ | $\frac{1}{2}\|\mu\|_2^2$ | $\frac{1}{2}\|\mu - \mu'\|_2^2$ |
| $\exp(\theta)$ | $\langle \mu, \log \mu \rangle - \mu$ | $\mu \log \frac{\mu}{\mu'}$ |
| $\log(1 + \exp(\theta))$ | $\eta \log \mu$ <br> $+(1-\mu)\log(1-\mu)$ | $(1-\mu)\log\left(\frac{1-\mu}{1-\mu'}\right)$ <br> $+\mu \log \frac{\mu}{\mu'}$ |

# Manifolds in primal and dual co-ordinates

$B_G(\cdot,\cdot)$ induces a Riemannian manifold $(\Theta, \nabla^2 G)$ in the primal co-ordinates.

$\Phi$ be the image of $\Theta$ under the continuous map $g = \nabla G$.
$B_H : \Phi \times \Phi \to \mathbb{R}^+$ induces the same Riemannian manifold $(\Phi, \nabla^2 H)$ under dual co-ordinates $\Phi$.

# Equivalence

**Theorem (Raskutti, Mukherjee)**

*The mirror descent step with Bregman divergence defined by $G$ applied to function $f$ in the space $\Theta$ is equivalent to the natural gradient step along Riemannian manifold $(\Phi, \nabla^2 H)$ in dual co-ordinates.*

# Consequences

Exponential family with density: $p(y \mid \theta) = h(y)\exp(\langle\theta, y\rangle - G(\theta))$.
Consider the following mirror descent step given $y_t$

$$\theta_{t+1} = \arg\min_{\theta}\left\{\left\langle\theta, \nabla_\theta B_G(\theta, h(y_t))|_{\theta=\theta_t}\right\rangle + \frac{1}{\alpha_t} B_G(\theta, \theta_t)\right\}.$$

In dual coordinates one would minimize

$$f_t(\mu; y_t) = -\log p(y_t \mid \mu) = B_H(y_t, \mu).$$

The natural gradient step is

$$\begin{aligned} \mu_{t+1} &= \mu_t - \alpha_t[\nabla^2 H(\mu_t)]^{-1}\nabla B_H(y_t, \mu_t), \\ &= \mu_{t+1} = \mu_t - \alpha_t(\mu_t - y_t), \end{aligned}$$

the curvature of the loss $B_H(y_t, \mu_t)$ matches the metric tensor $\nabla^2 H(\mu)$.

# Statistical efficiency

Given independent samples $Y_T = (y_1, ..., y_T)$ and a sequence of unbiased estimators $\widehat{\mu}_T$ is Fisher efficient if

$$\lim_{T\to\infty} \mathbb{E}_{Y_T}[(\widehat{\mu}_T - \mu)(\widehat{\mu}_T - \mu)^T] \to \frac{1}{T}\nabla^2 H,$$

where $\nabla^2 H$ is the inverse of the Fisher information matrix.

**Theorem (Raskutti, Mukherjee)**

*The mirror descent step applied to the log loss* (**??**) *with step-sizes* $\alpha_t = \frac{1}{t}$ *asymptotically achieves the Cramér-Rao lower bound.*

# Challenges

(1) Information geometry on mixture of manifolds.

(2) Proximity functions for functions over the Grassmannian.

(3) EM algorithms for mixtures.

# Acknowledgements

Funding:

- Center for Systems Biology at Duke
- NSF DMS and CCF
- DARPA
- AFOSR
- NIH