Advanced optimization methods
Mirror descent

Alexandr Katrutsa

MIPT department of applied mathematics and computer science

October 10, 2018

Plan for today[1]

- Uniform view on first order methods
- Mirror descent
- Bregman divergence
- Convergence analysis

[1] Pictures and some ideas are taken from this presentation

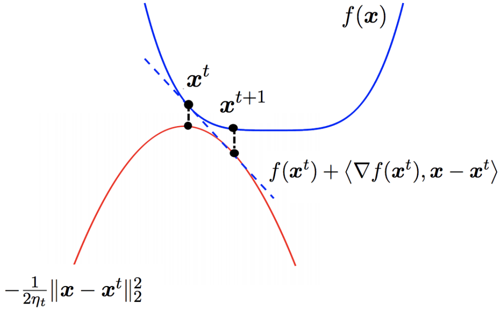
Problem statement

$$\min_{x \in G} f(x)$$

- f is smooth and convex
- G is convex and closed
- Condition number $\kappa = \frac{L}{\mu}$, where L is Lipschitz constant of gradient and μ is strong convexity constant

Projected gradient descent

$$
\begin{aligned}
x_{k+1} =& \pi_G(x_k - \alpha_k f'(x_k)) = \arg\min_{x \in G} \frac{1}{2}\|(x - x_k) + \alpha_k f'(x_k)\|_2^2 \\
=& \arg\min_{x \in G} \left\{ \underbrace{f(x_k) + \langle f'(x_k), x - x_k \rangle}_{\text{linear approximation}} + \underbrace{\frac{1}{2\alpha_k}\|x - x_k\|_2^2}_{\text{proximity term}} \right\}
\end{aligned}
$$



Use euclidean distance to measure discrepancy between f and FO approximation

Underlying problem geometry

- We believe that euclidean distance is good for local curvature estimation

Underlying problem geometry

- We believe that euclidean distance is good for local curvature estimation
- What is the main property of euclidean distance?

Underlying problem geometry

- We believe that euclidean distance is good for local curvature estimation
- What is the main property of euclidean distance?
- **Main issue**: local geometry might sometimes be highly inhomogeneous or even non-euclidean

Underlying problem geometry

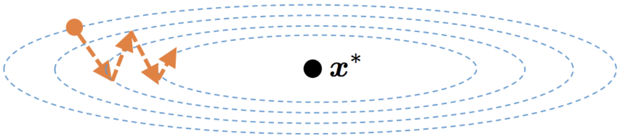
- We believe that euclidean distance is good for local curvature estimation
- What is the main property of euclidean distance?
- **Main issue**: local geometry might sometimes be highly inhomogeneous or even non-euclidean
- Can you give some examples?

Examples: quadratic programming

$$\min_{x} \frac{1}{2}(x - x^*)^\top A(x - x^*),$$

where $A \in \mathbb{S}^n_{++}$ and $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \gg 1$

- Gradient descent: $x_{k+1} = x_k - \alpha_k A(x_k - x^*)$ is slow, since convergence rate depends on κ



- It does not fit local curvature of f!

Examples: quadratic programming

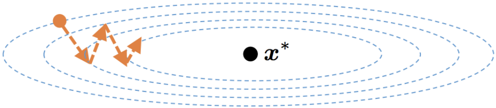
$$\min_x \frac{1}{2}(x - x^*)^\top A(x - x^*),$$

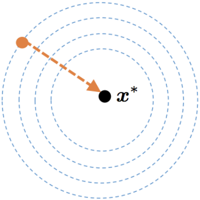
where $A \in \mathbb{S}^n_{++}$ and $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \gg 1$

- Rescaling gradient helps a lot

$$x_{k+1} = x_k - \alpha_k A^{-1} f'(x_k) = \underbrace{x_k - \alpha_k(x_k - x^*)}_{=x^* \text{ for } \alpha_k=1}$$

$$x_{k+1} = \arg\min_x \left\{ \langle f'(x_k), x - x_k\rangle + \underbrace{\frac{1}{2\alpha_k}(x - x_k)^\top A(x - x_k)}_{\text{proximity term}} \right\}$$





Examples: probability simplex

$$\min_{x \in \Delta} f(x),$$

where $\Delta = \{x \in \mathbb{R}^n_+ \mid x_1 + \ldots + x_n = 1\}$

- Euclidean distance is not appropriate to measure distance between probability vectors
- Different probability divergence metrics are better
- KL divergence
$$D_{KL}(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$$
- Total variation distance
- χ^2 divergence

Mirror descent: main idea

Nemirovsky & Yudin, 1983

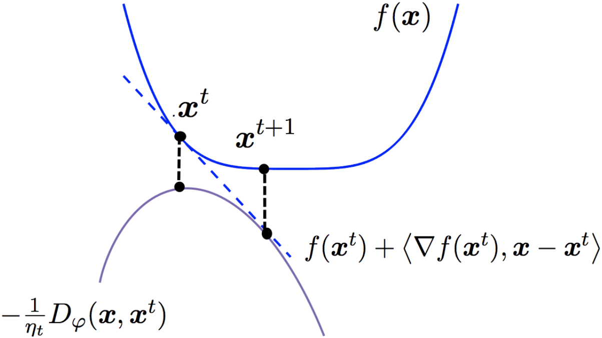
Adjust gradient updates to fit problem geometry

Mirror descent: formalism

Replace euclidean distance with distance-like function D_φ

$$x_{k+1} = \underset{x \in G}{\arg\min} \left\{ f(x_k) + \langle f'(x_k), x - x_k \rangle + \underbrace{\frac{1}{\alpha_k} D_\varphi(x, x_k)}_{\text{Bregman divergence}} \right\},$$

where $D_\varphi(x, z) = \varphi(x) - \varphi(z) - \langle \varphi'(z), x - z \rangle$ for convex and differentiable φ



Bregman divergence

Definition

Let φ be strictly convex and differentiable on G then for any $x, z \in G$

$$D_\varphi(x, z) = \varphi(x) - \varphi(z) - \langle \varphi'(z), x - z \rangle$$

- Similar to squared euclidean distance
- Locally quadratic measure:

$$D_\varphi(x, z) = (x - z)^\top \varphi''(\xi)(x - z)$$

for some ξ

How to choose Bregman divergence?

- Fit local curvature of f
- Use geometry of feasible set G
- Inexpensive computation of Bregman projection

Examples

- Squared Mahalanobis distance, $A \in \mathbb{S}^n_{++}$

$$\varphi(x) = \frac{1}{2} x^\top A x, \quad D_\varphi(x, z) = \frac{1}{2}(x - z)^\top A(x - z)$$

$$\text{MD: } x_{k+1} = x_k - \alpha_k A^{-1} f'(x_k)$$

- KL divergence for $G = \Delta$

$$\varphi(x) = \sum_i x_i \log x_i, \quad D_\varphi(x, z) = \sum_i x_i \log \frac{x_i}{z_i}$$

$$\text{MD: } x^i_{k+1} = \frac{x^i_k \exp(-\alpha_k [f'(x_k)]_i)}{\sum_{j=1}^n x^j_k \exp(-\alpha_k [f'(x_k)]_j)}$$

Also known as exponential gradient method

Some more cases

Table is from this paper

Function Name	$\varphi(x)$	$\operatorname{dom}\varphi$	$D_\varphi(x;y)$
Squared norm	$\frac{1}{2}x^2$	$(-\infty,+\infty)$	$\frac{1}{2}(x-y)^2$
Shannon entropy	$x\log x - x$	$[0,+\infty)$	$x\log\frac{x}{y} - x + y$
Bit entropy	$x\log x + (1-x)\log(1-x)$	$[0,1]$	$x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$
Burg entropy	$-\log x$	$(0,+\infty)$	$\frac{x}{y} - \log\frac{x}{y} - 1$
Hellinger	$-\sqrt{1-x^2}$	$[-1,1]$	$(1-xy)(1-y^2)^{-1/2} - (1-x^2)^{1/2}$
ℓ_p quasi-norm	$-x^p \qquad (0<p<1)$	$[0,+\infty)$	$-x^p + p\,xy^{p-1} - (p-1)\,y^p$
ℓ_p norm	$\|x\|^p \qquad (1<p<\infty)$	$(-\infty,+\infty)$	$\|x\|^p - p\,x\operatorname{sgn} y\,\|y\|^{p-1} + (p-1)\,\|y\|^p$
Exponential	$\exp x$	$(-\infty,+\infty)$	$\exp x - (x-y+1)\exp y$
Inverse	$1/x$	$(0,+\infty)$	$1/x + x/y^2 - 2/y$

Basic properties of Bregman divergence

Definition

Let φ be μ-strongly convex w.r.t. some norm in the domain X if

$$\varphi(x) \geq \varphi(y) + \langle \varphi'(y), x - y \rangle + \frac{\mu}{2} \|x - y\|$$

for every $x, y \in X$

Basic properties of Bregman divergence

Definition

Let φ be μ-strongly convex w.r.t. some norm in the domain X if

$$\varphi(x) \geq \varphi(y) + \langle \varphi'(y), x - y \rangle + \frac{\mu}{2}\|x - y\|$$

for every $x, y \in X$

Let $\varphi : G \to \mathbb{R}$ be μ-strongly convex and differentiable in G, then

Basic properties of Bregman divergence

Definition

Let φ be μ-strongly convex w.r.t. some norm in the domain X if

$$\varphi(x) \geq \varphi(y) + \langle \varphi'(y), x - y \rangle + \frac{\mu}{2}\|x - y\|$$

for every $x, y \in X$

Let $\varphi : G \to \mathbb{R}$ be μ-strongly convex and differentiable in G, then

- non-negativity: $D_\varphi(x, z) \geq 0$ and $D_\varphi(x, z) = 0$ iff $x = z$

Basic properties of Bregman divergence

Definition

Let φ be μ-strongly convex w.r.t. some norm in the domain X if

$$\varphi(x) \geq \varphi(y) + \langle \varphi'(y), x - y \rangle + \frac{\mu}{2}\|x - y\|$$

for every $x, y \in X$

Let $\varphi : G \to \mathbb{R}$ be μ-strongly convex and differentiable in G, then

- non-negativity: $D_\varphi(x, z) \geq 0$ and $D_\varphi(x, z) = 0$ iff $x = z$
- convexity: $D_\varphi(x, z)$ is convex in x, but not necessarily convex in z

Basic properties of Bregman divergence

Definition

Let φ be μ-strongly convex w.r.t. some norm in the domain X if

$$\varphi(x) \geq \varphi(y) + \langle \varphi'(y), x - y \rangle + \frac{\mu}{2}\|x - y\|$$

for every $x, y \in X$

Let $\varphi : G \to \mathbb{R}$ be μ-strongly convex and differentiable in G, then

- non-negativity: $D_\varphi(x, z) \geq 0$ and $D_\varphi(x, z) = 0$ iff $x = z$
- convexity: $D_\varphi(x, z)$ is convex in x, but not necessarily convex in z
- non-symmetric: in general $D_\varphi(x, z) \neq D_\varphi(z, x)$

Basic properties of Bregman divergence

Definition

Let φ be μ-strongly convex w.r.t. some norm in the domain X if

$$\varphi(x) \geq \varphi(y) + \langle \varphi'(y), x - y \rangle + \frac{\mu}{2}\|x - y\|$$

for every $x, y \in X$

Let $\varphi : G \to \mathbb{R}$ be μ-strongly convex and differentiable in G, then

- non-negativity: $D_\varphi(x, z) \geq 0$ and $D_\varphi(x, z) = 0$ iff $x = z$
- convexity: $D_\varphi(x, z)$ is convex in x, but not necessarily convex in z
- non-symmetric: in general $D_\varphi(x, z) \neq D_\varphi(z, x)$
- gradient: $(D_\varphi(x, z))'_x = \varphi'(x) - \varphi'(z)$

Three-point lemma

Lemma

For any three points x, y, z:

$$D_\varphi(x, z) = D_\varphi(x, y) + D_\varphi(y, z) - \langle \varphi'(z) - \varphi'(y), x - y \rangle$$

Proof on the blackboard

Q: what is the name of this lemma in the euclidean case?

Bregman projection

Definition

Given point x, then Bregman projection of x onto G is the following

$$\pi_{G,\varphi}(x) = \arg\min_{z \in G} D_{\varphi}(z, x)$$

We need fast method to find $\pi_{G,\varphi}$

Generalized Pythagorean theorem

Theorem

If $x_{G,\varphi} = \pi_{G,\varphi}(x)$, then

$$D_\varphi(z, x) \geq D_\varphi(z, x_{G,\varphi}) + D_\varphi(x_{G,\varphi}, x), \quad \forall z \in G$$

Proof on the blackboard

Why this descent is "mirror"?

- Rewrite original sub-problem with Bregman divergence

$$x_{k+1} = \arg\min_{x \in G} \left\{ \langle f'(x_k), x - x_k \rangle + \frac{1}{\alpha_k} D_\varphi(x, x_k) \right\}$$

- Optimality condition

$$0 \in N_G(x_{k+1}) + \alpha_k f'(x_k) + (\varphi'(x_{k+1}) - \varphi'(x_k))$$

- Bregman projection form

$$\varphi'(y_{k+1}) = \varphi'(x_k) - \alpha_k f'(x_k)$$
$$x_{k+1} = \arg\min_{x \in G} D_\varphi(x, y_{k+1})$$

Mirror between dual and primal spaces

Assume $G = \mathbb{R}^n$, then

$$x_{k+1} = y_{k+1} = (\varphi')^{-1}(\varphi'(x_k) - \alpha_k f'(x_k))$$

Mirror between dual and primal spaces

Assume $G = \mathbb{R}^n$, then

$$x_{k+1} = y_{k+1} = (\varphi')^{-1}(\varphi'(x_k) - \alpha_k f'(x_k))$$

- "Vectors" and "gradients" are from different spaces in general Banach space case

Mirror between dual and primal spaces

Assume $G = \mathbb{R}^n$, then

$$x_{k+1} = y_{k+1} = (\varphi')^{-1}(\varphi'(x_k) - \alpha_k f'(x_k))$$

- "Vectors" and "gradients" are from different spaces in general Banach space case
- Gradient descent does not make any sense!

Mirror between dual and primal spaces

Assume $G = \mathbb{R}^n$, then

$$x_{k+1} = y_{k+1} = (\varphi')^{-1}(\varphi'(x_k) - \alpha_k f'(x_k))$$

- "Vectors" and "gradients" are from different spaces in general Banach space case
- Gradient descent does not make any sense!
- We do not have this problem in Hilbert space, because of Fréchet–Riesz theorem

Mirror between dual and primal spaces

Assume $G = \mathbb{R}^n$, then

$$x_{k+1} = y_{k+1} = (\varphi')^{-1}(\varphi'(x_k) - \alpha_k f'(x_k))$$

- "Vectors" and "gradients" are from different spaces in general Banach space case
- Gradient descent does not make any sense!
- We do not have this problem in Hilbert space, because of Fréchet–Riesz theorem
- So, the main insight from MD is

Mirror between dual and primal spaces

Assume $G = \mathbb{R}^n$, then

$$x_{k+1} = y_{k+1} = (\varphi')^{-1}(\varphi'(x_k) - \alpha_k f'(x_k))$$

- "Vectors" and "gradients" are from different spaces in general Banach space case
- Gradient descent does not make any sense!
- We do not have this problem in Hilbert space, because of Fréchet–Riesz theorem
- So, the main insight from MD is
  1. Map x_k to the dual space with gradient of function that induces Bregman divergence

Mirror between dual and primal spaces

Assume $G = \mathbb{R}^n$, then

$$x_{k+1} = y_{k+1} = (\varphi')^{-1}(\varphi'(x_k) - \alpha_k f'(x_k))$$

- "Vectors" and "gradients" are from different spaces in general Banach space case
- Gradient descent does not make any sense!
- We do not have this problem in Hilbert space, because of Fréchet–Riesz theorem
- So, the main insight from MD is
    1. Map x_k to the dual space with gradient of function that induces Bregman divergence
    2. Perform gradient step in dual space

Mirror between dual and primal spaces

Assume $G = \mathbb{R}^n$, then

$$x_{k+1} = y_{k+1} = (\varphi')^{-1}(\varphi'(x_k) - \alpha_k f'(x_k))$$

- "Vectors" and "gradients" are from different spaces in general Banach space case
- Gradient descent does not make any sense!
- We do not have this problem in Hilbert space, because of Fréchet–Riesz theorem
- So, the main insight from MD is
  1. Map x_k to the dual space with gradient of function that induces Bregman divergence
  2. Perform gradient step in dual space
  3. Project new point in primal space w.r.t. Bregman divergence proximity

Conjugacy and inversion

Lemma

$$(\varphi')^{-1} = (\varphi^*)'$$

Proof

- Assume $y = \varphi'(x)$
- By definition $\langle x, y \rangle = \varphi(x) + \varphi^*(y)$
- From convexity of φ: $\langle x, y \rangle = \varphi^{**}(x) + \varphi^*(y)$
- From definition follows $x = (\varphi^*)'(y)$
- Finally $x = (\varphi^*)'(y) = (\varphi^*)'(\varphi'(x))$

Then unconstrained MD can be written as

$$x_{k+1} = (\varphi^*)'(\varphi'(x_k) - \alpha_k f'(x_k))$$

Convergence analysis: assumptions

- Problem statement
$$\min_{x \in G} f(x)$$
- f is convex and Lipschitz continuous
- G is convex and closed
- φ is ρ-strongly convex w.r.t. $\| \cdot \|$
- $\|g\|_* \leq L_f$ for any $g \in \partial f$, any point x, where $\| \cdot \|_*$ is dual norm

Convergence analysis: main theorem

Theorem

Assume f is convex and L_f-continuous on G and let φ be ρ-strongly convex w.r.t. $\|\cdot\|$. Then

$$f_K^{best} - f^* \leq \frac{\sup_{x\in G} D_\varphi(x, x_0) + \frac{L_f^2}{2\rho}\sum_{k=0}^K \alpha_k^2}{\sum_{k=0}^K \alpha_k}$$

- If $\alpha_k = \frac{\sqrt{2R\rho}}{L_f}\frac{1}{\sqrt{k}}$, where $R = \sup_{x\in G} D_\varphi(x, x_0)$, then

$$f_K^{best} - f^* \leq \mathcal{O}\left(\frac{L_f\sqrt{R}}{\sqrt{\rho}}\frac{\log k}{\sqrt{k}}\right)$$

- $\log$-factor can be eliminate

Optimization over probability simplex with ℓ_2

Assume $G = \Delta$ and $x_0 = n^{-1}\mathbf{1}$

(1) Use euclidean proximity term: $\varphi(x) = \frac{1}{2}\|x\|_2^2$ – 1-strongly convex in $\|\cdot\|_2$. Then

$$\sup_{x \in G} D_\varphi(x, x_0) = \sup_{x \in \Delta} \frac{1}{2}\|x - n^{-1}\mathbf{1}\|_2^2 = \sup_{x \in \Delta} \frac{1}{2}\left(\|x\|_2^2 - \frac{1}{n}\right) \leq \frac{1}{2}$$

and

$$f_K^{best} - f^* \leq \mathcal{O}\left(L_{f,2}\frac{\log k}{\sqrt{k}}\right),$$

i.e. for all subgradients g: $\|g'\|_2 \leq L_{f,2}$

Optimization over probability simplex with ℓ_1

Assume $G = \Delta$ and $x_0 = n^{-1}\mathbf{1}$

(2) Use ℓ_1 proximity term: $\psi(x) = -\sum_{i=1}^n x_i \log x_i$ – 1-strongly convex in $\|\cdot\|_1$. Then

$$\sup_{x\in G} D_\psi(x, x_0) = \sup_{x\in\Delta} D_{KL}(x||x_0) = \sup_{x\in\Delta} \sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n x_i \log\frac{1}{n}$$
$$= \log n + \sum_{i=1}^n x_i \log x_i \leq \log n$$

and

$$f_K^{best} - f^* \leq \mathcal{O}\left(L_{f,\infty}\sqrt{\log n}\frac{\log k}{\sqrt{k}}\right)$$

i.e. for all subgradients g: $\|g'\|_\infty \leq L_{f,\infty}$

Optimization over probability simplex: comparison

Ignore log-terms and compare

- Euclidean: $\mathcal{O}\left(\frac{L_{f,2}}{\sqrt{k}}\right)$
- D_{KL}: $\mathcal{O}\left(\frac{L_{f,\infty}}{\sqrt{k}}\right)$
- Equivalence norm

$$\|g\|_\infty \leq \|g\|_2 \leq \sqrt{n}\|g\|_\infty$$

- Why D_{KL} is better:

$$\frac{1}{\sqrt{n}} \leq \frac{L_{f,\infty}}{L_{f,2}} \leq 1$$

Highlights

- It is important to fit local geometry

Highlights

- It is important to fit local geometry
- Bregman divergence is effective tool to tune distance function for your problem

Highlights

- It is important to fit local geometry
- Bregman divergence is effective tool to tune distance function for your problem
- Idea from functional analysis is helpful here

Highlights

- It is important to fit local geometry
- Bregman divergence is effective tool to tune distance function for your problem
- Idea from functional analysis is helpful here
- Mirror descent separates steps in primal and dual spaces

Highlights

- It is important to fit local geometry
- Bregman divergence is effective tool to tune distance function for your problem
- Idea from functional analysis is helpful here
- Mirror descent separates steps in primal and dual spaces
- Experiments will be in the next class...