Introduction to mirror descent

Alexandr Katrutsa

April 15, 2021

Problem statement

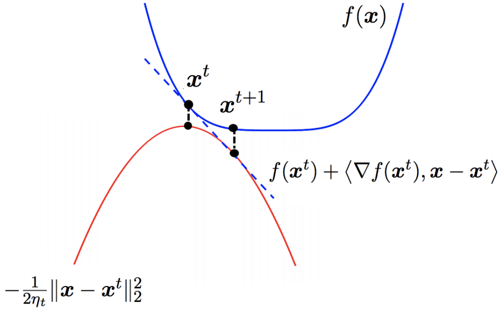
$$\min_{x \in G} f(x)$$

- 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

$$x_{k+1} = \pi_G(x_k - \alpha_k f'(x_k)) = \arg\min_{x \in G} \frac{1}{2}\|(x - x_k) + \alpha_k f'(x_k)\|_2^2$$

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



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?

Mirror descent: main idea

Nemirovsky & Yudin, 1983

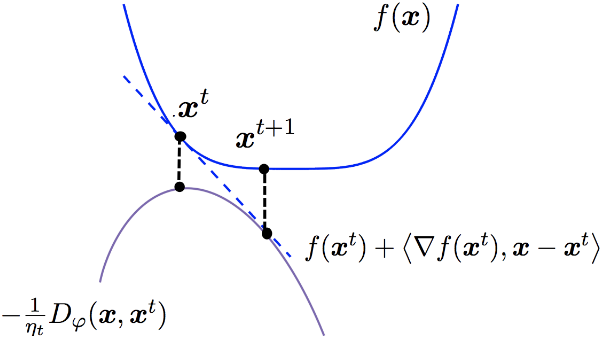
Adjust gradient updates to fit problem geometry

Mirror descent: formalism

Replace euclidean distance with distance-like function D_φ

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

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	$-x^p \quad (0<p<1)$	$[0,+\infty)$	$-x^p + p\,x y^{p-1} - (p-1)\,y^p$
ℓ_p norm	$\lvert x\rvert^p \quad (1<p<\infty)$	$(-\infty,+\infty)$	$\lvert x\rvert^p - p\,x\operatorname{sgn} y\,\lvert y\rvert^{p-1} + (p-1)\,\lvert y\rvert^p$
Exponential	$\exp x$	$(-\infty,+\infty)$	$\exp x - (x-y+1)\exp y$
Inverse	$1/x$	$(0,+\infty)$	$1/x + x/y^2 - 2/y$

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

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

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

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

and

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

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

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

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

$$= \log n + \sum_{i=1}^n x_i \log x_i \le \log n$$

and

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

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

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