Convex Optimization

Part 3: Mirror descent

Namhoon Lee

POSTECH

17 Oct 2022

Admin

Midterm

- result: 52.9 (avg) / 15.2 (std)
- check with TAs during office hours this week if you want

On dimension independent results

Consider constrained minimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \ & f(x) \\ \text{s.t. } & x \in \mathbb{C} \end{aligned}$$

For f G-Lipschitz, projected subgradient method with diminishing step size satisfies

$$f(x_{t,\text{best}}) - f(x^*) \leq \frac{RG}{\sqrt{T}}$$

- the bound has no dependence on n ("dimension free")
- in fact G is w.r.t. $\|\cdot\|_2$ and can be dimension dependent
- mirror descent aims to improve based upon this point

Gradient descent

Gradient descent as finding minimizer of function approximation

$$
\begin{aligned}
x^{+} &= \arg\min_{u} \; f(x) + \nabla f(x)^{\top}(u - x) + \frac{1}{2\eta}\|u - x\|_2^2 \\
&= \arg\min_{u} \; \eta \nabla f(x)^{\top} u + \underbrace{\frac{1}{2}\|u - x\|_2^2}_{\text{prox term}}
\end{aligned}
$$

- find u while staying close to x as measured in the Euclidean distance
- different distance measure (or geometry) gives a rise to mirror descent

Proximal gradient method

Proximal gradient method for minimizing composite function $f(x) = g(x) + h(x)$

$$\begin{aligned} x_{t+1} &= \text{prox}_{\eta h}(x_t - \eta \nabla g(x_t)) \\ &= \arg\min_u \left(h(u) + g(x_t) + \nabla g(x_t)^\top (u - x_t) + \frac{1}{2\eta} \|u - x_t\|_2^2 \right) \end{aligned}$$

Quadratic term represents

- a penalty that forces x_{t+1} to be close to x_t, where linearization of g is accurate
- an approximation of the erorr term in the linearization of g at x_t

Generalized proximal gradient method

Replace $\frac{1}{2}\|u - x\|_2^2$ with a generalized distance $D(u, x)$

$$x_{t+1} = \arg\min_u \left(h(u) + g(x_t) + \nabla g(x_t)^\top (u - x_t) + \frac{1}{\eta} D(u, x_t) \right)$$

Potential benefits

- "pre-conditioning": use a more accurate model of $g(u)$ around x, ideally

$$\frac{1}{\eta} D(u, x_t) \approx g(u) - g(x_t) - \nabla g(x_t)^\top (u - x_t)$$

- make the generalized proximal mapping (minimizer u) easier to compute

Bregman distance

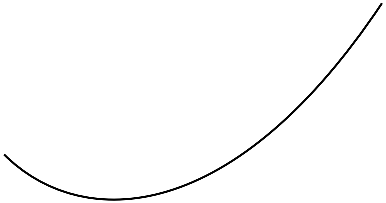
Definition

$$D_\phi(x, y) = \phi(x) - \phi(y) - \nabla\phi(y)^\top(x - y)$$

- ϕ is convex and continuously differentiable on $\mathrm{int}(\mathrm{dom}\,\phi)$
- ϕ is called kernel function or distance-generating function

Read "distance between x and y as measured by function ϕ" or "divergence from x to y with respect to function ϕ"

Illustration



Immediate properties

Bregman distance

$$D_\phi(x, y) = \phi(x) - \phi(y) - \nabla\phi(y)^\top(x - y)$$

- $D_\phi(x, y)$ is convex in x for fixed y
- $D_\phi(x, y) \geq 0$ with equality if $x = y$
- if ϕ is strictly convex, then $D_\phi(x, y) = 0$ only if $x = y$
- $D_\phi(x, y) \neq D_\phi(y, x)$ in general

to emphasize lack of symmetry, D is also called a directed distance or divergence

Examples

ϕ squared 2-norm

$$\phi(x) = \frac{1}{2}\|x\|_2^2$$

Bregman distance

$$D_\phi(x, y) = \frac{1}{2}\|x - y\|_2^2$$

i.e., squared Euclidean distance

- reduces to gradient descent

Examples

ϕ general quadratic

$$\phi(x) = \frac{1}{2} x^\top A x$$

where A is symmetric positive (semi)definite

Bregman distance

$$D_\phi(x, y) = \frac{1}{2}(x - y)^\top A (x - y)$$

i.e., general quadratic kernel

- leads to pre-conditioning

Examples

ϕ unnormalized negative entropy

$$\phi(x) = \sum_{i=1}^{n} x_i \log x_i - x_i$$

Bregman distance

$$D_\phi(x, y) = \sum_{i=1}^{n} x_i \log \frac{x_i}{y_i} - x_i + y_i$$

i.e., unnormalized relative entropy or KL divergence

Three-point identity

For all $x \in \operatorname{dom} \phi$ and $y, z \in \operatorname{int}(\operatorname{dom} \phi)$

$$D_\phi(x, z) = D_\phi(x, y) + D_\phi(y, z) + (\phi(y) - \phi(z))^\top (x - y)$$

- proof is done straightforward by substituting the definition of D_ϕ

Strongly convex kernel

We will sometimes assume that ϕ is strongly convex

$$\phi(x) \geq \phi(y) + \nabla\phi(y)^\top(x-y) + \frac{\alpha}{2}\|x-y\|^2$$

- $\alpha > 0$ is strong convexity constant of ϕ for the norm $\|\cdot\|$
- for twice differentiable ϕ, this is equivalent to

$$\nabla^2\phi(x) \succeq \alpha I$$

- strong convexity of ϕ implies that

$$D_\phi(x,y) = \phi(x) - \phi(y) - \nabla\phi(y)^\top(x-y) \geq \frac{\alpha}{2}\|x-y\|^2$$

Regularization with Bregman distance

For given $y \in \mathrm{int}(\mathrm{dom}\,\phi)$ and convex f, consider

$$\min f(x) + D_\phi(x, y)$$

- equivalently, minimize $f(x) + \phi(x) - \nabla\phi(y)^\top x$
- feasible set is $\mathrm{dom}\, f \cap \mathrm{dom}\,\phi$

Optimality condition: $\hat{x} \in \mathrm{dom}\, f \cap \mathrm{int}(\mathrm{dom}\,\phi)$ is optimal if and only if

$$\nabla\phi(y) - \nabla\phi(\hat{x}) \in \partial f(\hat{x})$$

Mirror descent

$$\min \; f(x)$$
$$\text{s.t. } x \in \mathbb{C}$$

- f is a convex function, $\mathbb{C}$ is a convex subset of $\operatorname{dom} f$
- we assume f is subdifferentiable on C

Algorithm: start with x_1 and repeat

$$x_{t+1} = \operatorname*{arg\,min}_{x \in \mathbb{C}} \eta g_t^\top x + D_\phi(x, x_t) \;, \quad t = 1, 2, \ldots$$

where $g_t \in \partial f(x_t)$

Mirror descent with quadratic kernel

$$x_{t+1} = \arg\min_{x \in \mathbb{C}} \eta g_t^\top x + D_\phi(x, x_t)$$

for $D_\phi(x, y) = \frac{1}{2}\|x - y\|_2^2$, this is the projected subgradient method

$$\begin{aligned} x_{t+1} &= \arg\min_{x \in \mathbb{C}} \eta g_t^\top x + \frac{1}{2}\|x - x_t\|_2^2 \\ &= \arg\min_{x \in \mathbb{C}} \frac{1}{2}\|x - x_t + \eta g_t\|_2^2 \\ &= \mathrm{P}_{\mathbb{C}}(x_t - \eta g_t) \end{aligned}$$

Mirror map view

Mirror descent (without constraint for simplicity)

$$x_{t+1} = \arg\min_{x} \eta \nabla f(x_t)^\top x + D_\phi(x, x_t)$$

Applying optimality condition

$$\nabla \phi(x_{t+1}) = \nabla \phi(x_t) - \eta \nabla f(x_t)$$

Taking $\nabla \phi$ as an operator (or mapping)

$$x_{t+1} = (\nabla \phi)^{-1}(\nabla \phi(x_t) - \eta \nabla f(x_t))$$

With Bregman projection

$$y_{t+1} = (\nabla\phi)^{-1}(\nabla\phi(x_t) - \eta\nabla f(x_t)) \ , \qquad x_{t+1} = \arg\min_{x\in\mathcal{X}} D_\phi(x, y_{t+1})$$

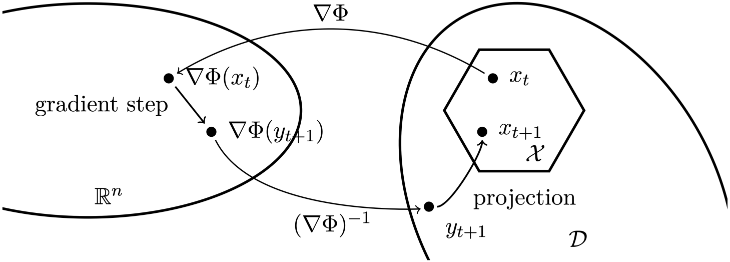


Figure: Illustration of mirror descent; figure from Bubeck

View ϕ as "mirror map" and $\nabla\phi(x)$ as the point mapped from primal to dual space (Nemirovskij and Yudin 1983)

Running examples

for $\phi(x) = \frac{1}{2}\|x\|_2^2$, $\nabla\phi(x) = x$, so we get

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

- gradient descent

for $\phi(x) = \sum_{i=1}^n x_i \log x_i - x_i$, $\nabla\phi(x) = (\log x_1, \ldots, \log x_n)$, so we get

$$(x_{t+1})_i = (x_t)_i \exp\big(-\eta(\nabla f(x_t))_i\big)$$

- Hedge algorithm (with normalization step for constrained case)

Dual norm

Definition (Dual norm)

Let $\|\cdot\|$ be some norm. Its dual norm is defined as

$$\|x\|_* = \sup_{\|y\|\leq 1} y^\top x$$

- dual norm of 2-norm is 2-norm itself ("self-dual")
- dual norm of p-norm is q-norm where $1/p + 1/q = 1$
- dual of dual norm is the original norm itself ($(\|\cdot\|_*)_* = \|\cdot\|$)

Cauchy-Schwarz for general norms

For $x, y \in \mathbb{R}^n$, we have

$$\langle x, y \rangle \leq \|x\| \|y\|_*$$

Proof.

Dividing both sides by $\|x\|$ yields $\langle x/\|x\|, y \rangle \leq \|y\|_*$. The inequality holds by definition of dual norm and by noting that $\|x/\|x\|\| = 1$ for $\|x\| \neq 0$. □

Convergence analysis

Theorem

Let f be convex and L-Lipschitz w.r.t. $\|\cdot\|$. Let ϕ be ρ-strongly convex with respect to $\|\cdot\|$. Mirror descent with $\eta = \frac{R}{L}\sqrt{\frac{2\rho}{T}}$ and $R^2 \geq D_\phi(x, x^*)$ satisfies

$$f\left(\frac{1}{T}\sum_{i=1}^{T} x_t\right) - f(x^*) \leq RL\sqrt{\frac{2}{\rho T}}$$

- $1/\sqrt{T}$ same dependence on T for subgradient
- L is w.r.t. $\|\cdot\|$ not $\|\cdot\|_2$

Proof.

We start with convexity of f and some algebraic manipulation

$$\begin{aligned}\eta(f(x_t) - f(x^*)) &\leq \eta(g_t^\top (x_t - x^*)) \\ &= \underbrace{(\nabla\phi(x_t) - \nabla\phi(x_{t+1}) - \eta g_t)^\top (x^* - x_{t+1})}_{A} \\ &\quad + \underbrace{(\nabla\phi(x_{t+1}) - \nabla\phi(x_t))^\top (x^* - x_{t+1})}_{B} + \underbrace{\eta g_t^\top (x_t - x_{t+1})}_{C}\end{aligned}$$

We will bound each term A, B, C

first prove $A \leq 0$

$$A = (\nabla\phi(x_t) - \nabla\phi(x_{t+1}) - \eta g_t)^\top (x^* - x_{t+1})$$

recall mirror descent

$$x_{t+1} = \arg\min_{x \in \mathbb{C}} \eta g_t^\top x + D_\phi(x, x_t)$$

recall optimality condition for convex optimization with set constraint

$$0 \in \eta g_t + \nabla\phi(x_{t+1}) - \nabla\phi(x_t) + \mathcal{N}_{\mathbb{C}}(x_{t+1})$$

by the definition of normal cone

$$(\eta g_t + \nabla\phi(x_{t+1}) - \nabla\phi(x_t))^\top (x - x_{t+1}) \geq 0 \quad \forall x \in \mathbb{C}$$

therefore

$$A \leq 0$$

we can express B as follows by the definition of Bregman distance

$$\begin{aligned} B &= (\nabla\phi(x_{t+1}) - \nabla\phi(x_t))^\top (x^* - x_{t+1}) \\ &= D_\phi(x^*, x_t) - D_\phi(x_{t+1}, x_t) - D_\phi(x^*, x_{t+1}) \end{aligned}$$

next bound C

$$C = \eta g_t^\top (x_t - x_{t+1}) \leq \frac{1}{2\rho}\eta^2 \|g_t\|_*^2 + \frac{\rho}{2}\|x_t - x_{t+1}\|^2$$

where we use Hölder's inequality

$$u^\top v \leq \frac{1}{2\alpha}\|u\|_*^2 + \frac{\alpha}{2}\|v\|^2$$

- generalization of completing square to non-Euclidean geometry put A B C together

$$\eta(f(x_t) - f(x^*)) \le D_\phi(x^*, x_t) - D_\phi(x_{t+1}, x_t) - D_\phi(x^*, x_{t+1}) + \frac{1}{2\rho}\eta^2\|g_t\|_*^2 + \frac{\rho}{2}\|x_t - x_{t+1}\|^2$$

use strong convexity of ϕ (*i.e.* $D_\phi(x_{t+1}, x_t) \ge (\rho/2)\|x_{t+1} - x_t\|^2$)

$$\eta(f(x_t) - f(x^*)) \le \underbrace{D_\phi(x^*, x_t) - D_\phi(x^*, x_{t+1})}_{\text{telescoping}} + \frac{1}{2\rho}\eta^2\|g_t\|_*^2$$

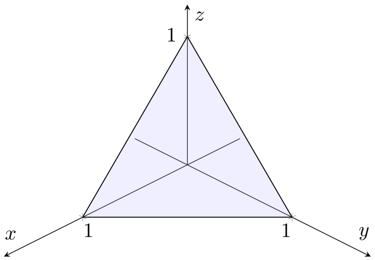
sum for T iterations and up to trivial computation

$$f\left(\frac{1}{T}\sum_{i=1}^{T} x_t\right) - f(x^*) \le \frac{D_\phi(x^*, x_1)}{\eta T} + \frac{\eta L^2}{2\rho}$$

□

Simplex setup

For minimizing on simplex constraint $\mathbb{C} = \Delta_n = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$, mirror descent with ϕ the negative entropy achieves a rate of convergence of order $\sqrt{\frac{\log n}{T}}$ whereas subgradient method only achieves $\sqrt{\frac{n}{t}}$.
For ϕ the negative entropy, we can show that ϕ is 1-strongly convex w.r.t. $\|\cdot\|_1$ (Pinsker's inequality).



Any questions?

References I

Nemirovskij, Arkadij Semenovič and David Borisovich Yudin (1983). "Problem complexity and method efficiency in optimization". In: *Wiley-Interscience*.