

lec. 20 (12/01/2022)

(The Last!!)

We assume that ψ is differentiable & strictly convex, $\psi: \text{dom}(\psi) \mapsto \mathbb{R}$ such that the constraint set satisfies:

- $\mathcal{X} \subset \text{closure}(\text{dom}(\psi))$,

- $\text{range}(\nabla \psi) = \mathbb{R}^n$,

- and $\|\nabla \psi\|_2 \rightarrow +\infty$ as $\underline{x} \rightarrow \text{boundary of closure}(\text{dom}(\psi))$

Then we say $\psi(\cdot)$ is a mirror map.

Conceptually, what's going on?

Take \underline{x}_k $\xrightarrow{\text{map to}}$ dual space by taking the gradient/subgradient

↑
element of the
primal space

do
↓

update in gradient/dual
space

primal space
to obtain \underline{x}_{k+1}

←
come back

via projection
w.r.t. D_ψ

(Bregman projection)

Bregman divergence:

Given a mirror map Ψ , the associated Bregman divergence $D_\Psi : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}_{\geq 0}$ is defined as:

$$D_\Psi(\underline{x}, \underline{y}) := \Psi(\underline{x}) - \Psi(\underline{y}) - \langle \nabla \Psi(\underline{y}), \underline{x} - \underline{y} \rangle$$
$$\forall \underline{x}, \underline{y} \in \mathcal{X}.$$

Very clean geometric interpretation:
 $D_\Psi(\underline{x}, \underline{y})$ quantifies the amount by which a strictly convex function lies above its tangent hyperplane

Example:

• $\psi(\underline{x}) = \|\underline{x}\|_2^2$, $\mathcal{X} \subset \mathbb{R}^n$ (closed)

$$\begin{aligned} D_\psi(\underline{x}, \underline{y}) &= \|\underline{x}\|_2^2 - \|\underline{y}\|_2^2 - \langle 2\underline{y}, \underline{x} - \underline{y} \rangle \\ &= \|\underline{x}\|_2^2 - \|\underline{y}\|_2^2 + 2\|\underline{y}\|_2^2 - \langle 2\underline{y}, \underline{x} \rangle \\ &= \|\underline{x}\|_2^2 + \|\underline{y}\|_2^2 - 2\langle \underline{y}, \underline{x} \rangle \\ &= \|\underline{x} - \underline{y}\|_2^2 \end{aligned}$$

• $\psi(\underline{x}) = \sum_{i=1}^n x_i \log x_i$, $\mathcal{X} \equiv \Delta^{n-1} = \{\underline{x} \in \mathbb{R}_{\geq 0}^n \mid \sum_{i=1}^n x_i = 1\}$

$$D_\psi(\underline{x}, \underline{y}) = \sum_{i=1}^n x_i \log\left(\frac{x_i}{y_i}\right), \text{ Kullback-Leibler divergence}$$

In general, $D_\psi(\underline{x}, \underline{y})$ as some kind of distance squared
we think about.

But $D_\psi(\underline{x}, \underline{y})$ may not be a metric,
in general

e.g., $\underbrace{D_\psi(\underline{x}, \underline{y}) \neq D_\psi(\underline{y}, \underline{x})}_{\text{maybe non-symmetric}}$

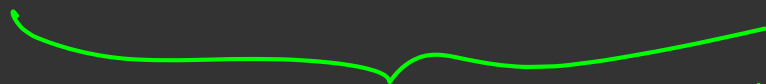
e.g., $D_\psi(\underline{x}, \underline{y})$ may not satisfy the
triangle inequality.

- When $\psi(\cdot) = \|\cdot\|_2^2$, Then $D_\psi(\underline{x}, \underline{y}) = \|\underline{x} - \underline{y}\|_2^2$
and mirror descent $\xrightarrow[\text{to}]{\text{reduces}}$ projected grad. descent

- When $\psi(\cdot) = \sum_{i=1}^n x_i \log x_i$ (neg. entropy)

$$\mathcal{X} = \Delta^{n-1}$$

$$D_\psi(\underline{x}, \underline{y}) = \sum_{i=1}^n x_i \log \left(\frac{x_i}{y_i} \right)$$



Kullback-Leibler divergence.

Then the mirror descent algorithm becomes:

$$\left\{ \begin{array}{l} \underline{y}_{k+1} = \underline{x}_k \odot \exp(-\eta_k \nabla f(\underline{x}_k)) \\ \underline{x}_{k+1} = \frac{\underline{y}_{k+1}}{\mathbf{1}^T \underline{y}_{k+1}} \end{array} \right\} \text{Kullback-Leibler projection} \\ \text{(see Lec. 16, p. 10-15)}$$

where \odot denotes elementwise vector-vector multiplication