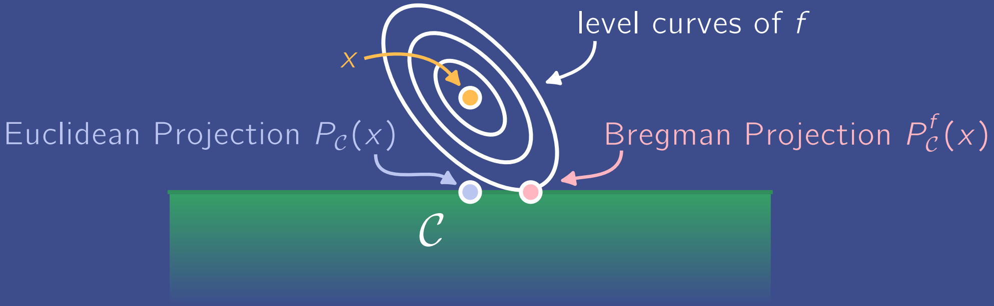
# Bregman Divergences

A natural way to measure closeness




Howard Heaton

Typal Academy

# Motivation

Optimization algorithms should respect problem geometry

Standard algorithms (*e.g.* proximal gradient) use Euclidean geometry

These slides describe how to use a function $f$ for various geometries

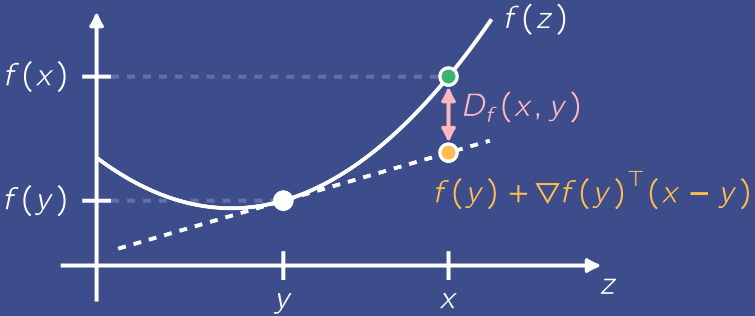
Note: We assume $f$ is strictly convex, proper, and differentiable in its domain

# Definition

The Bregman divergence $D_f$ associated with $f$ is given by

$$D_f(x, y) = f(x) - f(y) - \nabla f(y)^\top (x - y)$$

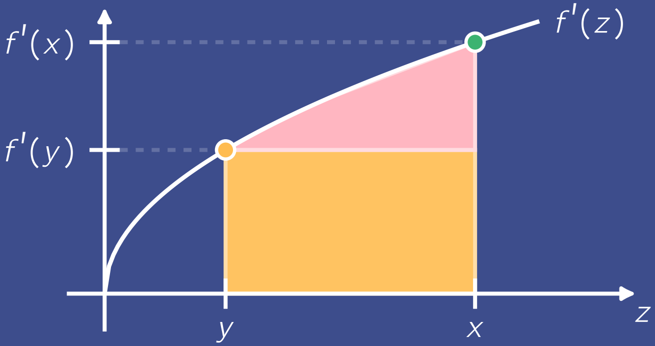
*i.e.* difference between function value at $x$ and a linear estimate from $y$ to $x$



# Area View of Bregman Divergence

In one dimension, this divergence can be represented as an area:

$$\text{Area}(\text{[pink region]}) = D_f(x, y) = \int_y^x f'(z)\,\mathrm{d}z - f'(y)(x - y)$$



# Key Properties

- Convexity: $D_f(x, y)$ is convex in $x$ when $y$ is fixed
- Positivity: $D_f(x, y) \geq 0$ with equality if and only if $x = y$
- Asymmetry: possible to have $D_f(x, y) \neq D_f(y, x)$

# Example – Ellipsoidal Norms

If $M$ is a positive definite matrix and $f(x) = \frac{1}{2}\|x\|_M^2 = \frac{1}{2}x^\top Mx$, then

$$\begin{aligned} D_f(x,y) &= \frac{1}{2}x^\top Mx - \frac{1}{2}y^\top My - (My)^\top(x-y) \\ &= \frac{1}{2}(x-y)^\top M(x-y) \\ &= \frac{1}{2}\|x-y\|_M^2 \end{aligned}$$

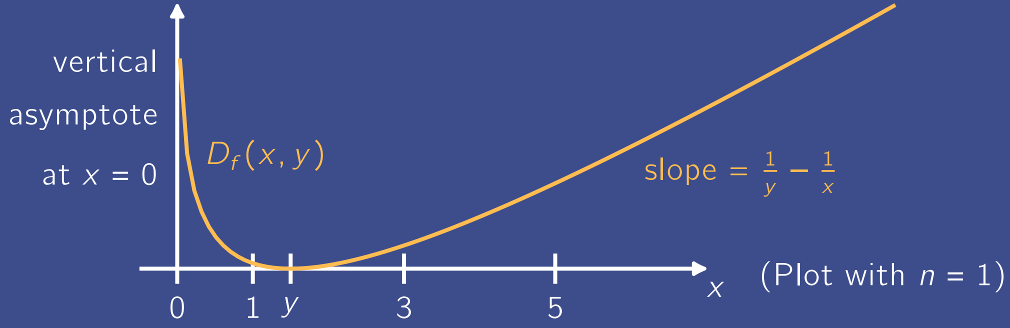
Special Case: If $M = \mathrm{I}$, Bregman divergence equals Euclidean distance squared,

$$i.e.\ D_f(x,y) = \tfrac{1}{2}\|x-y\|^2$$

# Example – Logarithmic Barrier

The logarithmic barrier uses $f\colon (0,\infty)^n \to \mathbb{R}$ given by

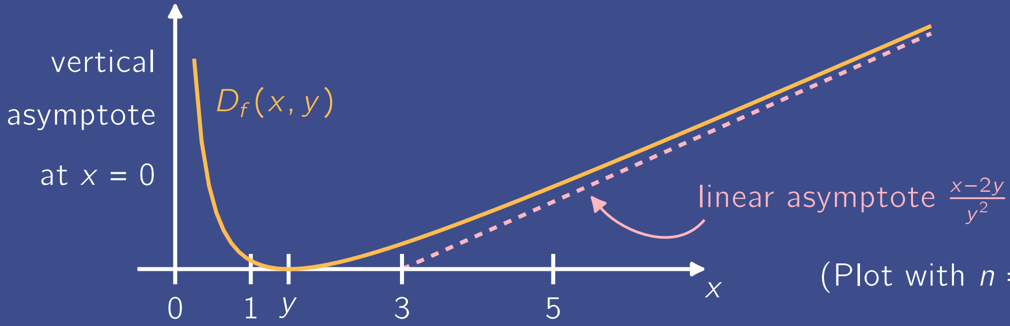
$$f(x) = -\sum_{i=1}^{n} \ln(x_i) \quad \Longrightarrow \quad D_f(x,y) = \sum_{i=1}^{n} \left( \frac{x_i}{y_i} - \ln\left(\frac{x_i}{y_i}\right) - 1 \right)$$



# Example – Inverse Barrier

The inverse barrier uses $f\colon (0,\infty)^n \to \mathbb{R}$ given by

$$f(x) = \sum_{i=1}^{n} \frac{1}{x_i} \quad \Longrightarrow \quad D_f(x,y) = \sum_{i=1}^{n} \frac{1}{x_i} + \frac{x_i - 2y_i}{y_i^2}$$




(Plot with $n = 1$)

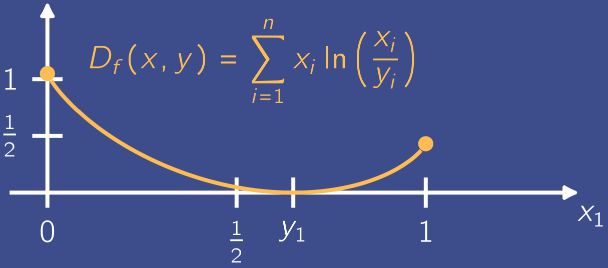
# Example – KL Divergence

The negative entropy function $f\colon (0,\infty)^n \to \mathbb{R}$ is given by

$$f(x) = \sum_{i=1}^{n} x_i \ln(x_i) \qquad \text{(Negative Entropy)}$$

If $x$ and $y$ are in the unit simplex with $y_i > 0$ for each $i$, then[†]

$$D_f(x, y) = \sum_{i=1}^{n} x_i \ln\left(\frac{x_i}{y_i}\right) \qquad \text{(KL Divergence)}$$



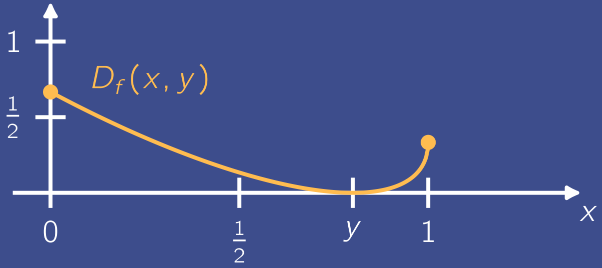

(Plot with $n = 2$)

[†]We adopt the convention $0 \ln(0) = 0$

# Example – Divergence with Square Root

Consider the divergence using $f\colon [0,1]^n \to \mathbb{R}$ given by

$$f(x) = -\sum_{i=1}^{n} \sqrt{1 - x_i^2} \qquad \Longrightarrow \qquad D_f(x, y) = \sum_{i=1}^{n} \left( -\sqrt{1 - x_i^2} + \frac{1 - x_i y_i}{\sqrt{1 - y_i^2}} \right)$$



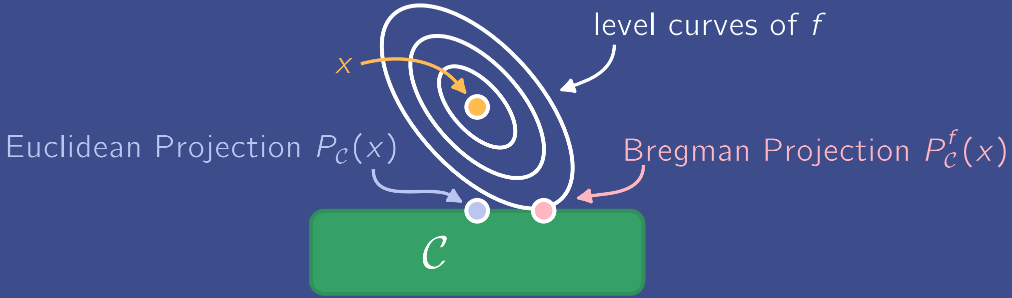

(Plot with $n = 1$)

# Bregman Projection

If $\mathcal{C}$ is a closed, convex, and nonempty set, then the Bregman projection

$$P^f_{\mathcal{C}}(x) = \operatorname*{argmin}_{z \in \mathcal{C}} D_f(z, x)$$

exists and is unique



# Example – Bregman Projection onto Hyperplane

For a scalar $\beta$, consider the hyperplane $\mathcal{H} = \{x : \sum_{i=1}^{n} x_i = \beta\}$

The standard Euclidean projection onto $\mathcal{H}$ is[†]

$$P_{\mathcal{C}}(x) = \operatorname*{argmin}_{z \in \mathcal{H}} \frac{1}{2}\|z - x\|^2 = x - \frac{\mathbf{1}^\top x - \beta}{n}\mathbf{1}$$

Picking $f$ to be negative entropy yields

$$P_{\mathcal{C}}^{f}(x) = \operatorname*{argmin}_{z \in \mathcal{H}} \sum_{i=1}^{n} z_i \ln\left(\frac{z_i}{x_i}\right) - z_i + x_i = \frac{\beta x}{\mathbf{1}^\top x}$$

[†]Euclidean distance uses $f(x) = \frac{1}{2}\|x\|^2$ and $\mathbf{1}$ is the vector of all ones

# Algorithm – Mirror Descent

Consider the constrained minimization problem

$$\min_{x \in \mathcal{C}} g(x)$$

Using step sizes $\alpha_k$, projected gradient updates take the form

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathcal{C}} g(x^k) + (x - x^k)^\top \nabla g(x^k) + \frac{1}{2\alpha_k} \|x - x^k\|^2$$

Mirror descent generalizes this to

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathcal{C}} g(x^k) + (x - x^k)^\top \nabla g(x^k) + \frac{1}{\alpha_k} D_f(x, x^k)$$

# Example – Mirror Descent on Nonnegative Orthant

Suppose the constraint set $\mathcal{C} = \{x : x_i \geq 0 \text{ for all } i\}$

Mirror descent with $f$ as negative entropy has updates of the form[†]

$$x^{k+1} = \operatorname*{argmin}_x x^\top \nabla g(x^k) + \frac{1}{\alpha_k} \cdot \sum_{i=1}^{n} x_i \ln\left(\frac{x_i}{x_i^k}\right) - x_i + x_i^k$$

which simplifies to

$$x^{k+1} = x^k \bullet \exp\left(-\alpha_k \nabla g(x^k)\right)$$

where $\bullet$ denotes element-wise multiplication

[†] A subscript $i$ is used to denote the $i$-th component of vectors

# Example – Mirror Descent on Simplex

Suppose the constraint set $\mathcal{C}$ is the unit simplex $\Delta_n$

Mirror descent with $f$ as negative entropy has updates of the form

$$w^k = \left( \sum_{j=1}^{n} x_j^k e^{-\alpha_k \nabla g(x^k)_j} \right)^{-1}$$

$$x_i^{k+1} = w^k x_i^k e^{-\alpha_k \nabla g(x^k)_i} \quad \text{for all } i = 1, 2, \ldots, n$$

The $w^k$ term ensures each $x^k$ is in the simplex $\Delta_n$

# Algorithm – Generalized Proximal Gradient

Consider the minimization of the sum of two convex functions:[†]

$$\min_x g(x) + h(x)$$

Proximal gradient updates take the form

$$x^{k+1} = \operatorname*{argmin}_x g(x) + h(x^k) + (x - x^k)^\top \nabla h(x^k) + \frac{1}{2\alpha_k}\|x - x^k\|^2$$

This can be generalized to

$$x^{k+1} = \operatorname*{argmin}_x g(x) + h(x^k) + (x - x^k)^\top \nabla h(x^k) + \frac{1}{\alpha_k} D_f(x, x^k)$$

[†] This is a generalization of the problem for mirror descent

# Convergence of Generalized Proximal Gradient

## Assumptions[†]

- $h$ is differentiable, $\mathrm{dom}(f) \subseteq \mathrm{dom}(h)$
- $L \cdot f(x) - h(x)$ is convex for some $L > 0$
- If $x^k \in \mathrm{int}(\mathrm{dom}(f))$, then $x^{k+1} \in \mathrm{int}(\mathrm{dom}(f))$
- Minimizer is obtained at $x^\star \in \mathrm{dom}(f)$ and $g(x^\star) + h(x^\star)$ is finite

Picking $\alpha_k = 1/L$ yields $g(x^k) + h(x^k) \leq g(x^\star) + h(x^\star) + \frac{L \cdot D_f(x^\star, x^1)}{k}$

[†]These are in addition to those previously stated

**Found this useful?**

➕ Follow for more

♻️ Repost to share with friends