

Mirror descent

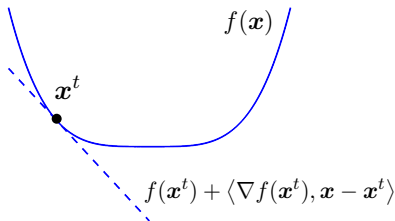
Ye Shi

ShanghaiTech University

Outline

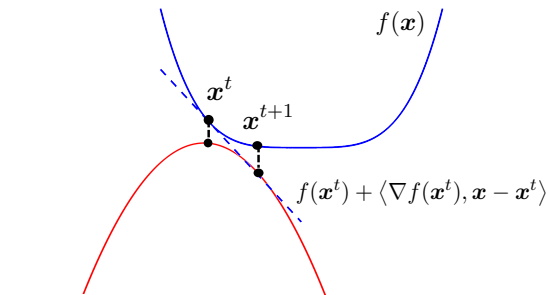
- Mirror descent
- Bregman divergence
- Alternative forms of mirror descent
- Convergence analysis

A proximal viewpoint of projected GD



$$x^{t+1} = \arg \min_{x \in \mathcal{C}} \left\{ \underbrace{f(x^t) + \langle \nabla f(x^t), x - x^t \rangle}_{\text{linear approximation}} + \frac{1}{2\eta_t} \|x - x^t\|_2^2 \right\}$$

A proximal viewpoint of projected GD



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

- the quadratic proximal term is used by GD to monitor the discrepancy between $f(\cdot)$ and its first-order approximation

Inhomogeneous / non-Euclidean geometry

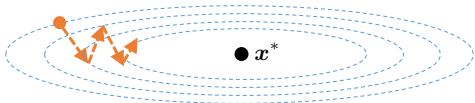
The quadratic proximity term is based on certain “prior belief”:

- the discrepancy between $f(\cdot)$ and its linear approximation is locally well approximated by the *homogeneous* penalty

$$\underbrace{(2\eta_t)^{-1} \|\mathbf{x} - \mathbf{x}^t\|_2^2}_{\text{squared Euclidean penalty}}$$

Issues: the local geometry might sometimes be highly *inhomogeneous*, or even *non-Euclidean*

Example: quadratic minimization

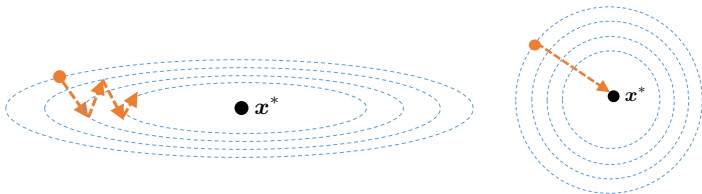


$$\text{minimize}_{x \in \mathbb{R}^n} \quad f(x) = \frac{1}{2}(x - x^*)^\top Q(x - x^*)$$

where $Q \succ 0$ is a diagonal matrix with large $\kappa = \frac{\max_i Q_{i,i}}{\min_i Q_{i,i}} \gg 1$

- gradient descent $x^{t+1} = x^t - \eta_t Q(x^t - x^*)$ is slow, since the iteration complexity is $O(\kappa \log \frac{1}{\varepsilon})$
- doesn't fit the curvature of $f(\cdot)$ well

Example: quadratic minimization



$$\text{minimize}_{x \in \mathbb{R}^n} \quad f(x) = \frac{1}{2}(x - x^*)^\top Q(x - x^*)$$

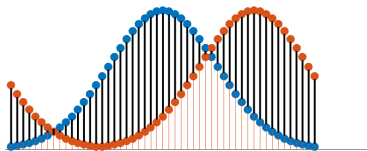
where $Q \succ 0$ is a diagonal matrix with large $\kappa = \frac{\max_i Q_{i,i}}{\min_i Q_{i,i}} \gg 1$

- one can significantly accelerate it by *rescaling* the gradient

$$x^{t+1} = x^t - \eta_t Q^{-1} \nabla f(x^t) = \underbrace{x^t - \eta_t (x^t - x^*)}_{\text{reaches } x^* \text{ in 1 iteration with } \eta_t=1}$$

$$\Leftrightarrow x^{t+1} = \arg \min_{x \in \mathbb{R}^n} \left\{ \langle \nabla f(x^t), x - x^t \rangle + \underbrace{\frac{1}{2\eta_t} (x - x^t)^\top Q (x - x^t)}_{\text{fits geometry better}} \right\}$$

Example: probability simplex



total-variation distance

$$\text{minimize}_{\mathbf{x} \in \Delta} f(\mathbf{x})$$

where $\Delta := \{\mathbf{x} \in \mathbb{R}_+^n \mid \mathbf{1}^\top \mathbf{x} = 1\}$ is probability simplex

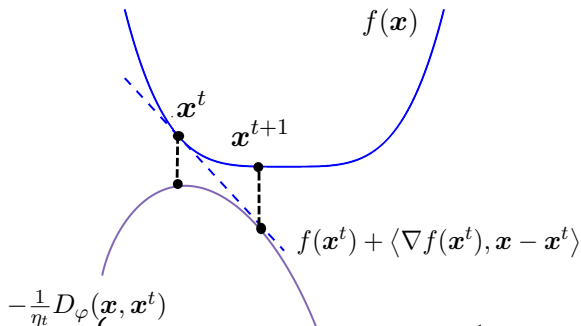
- Euclidean distance is in general not recommended for measuring the distance between probability vectors
- may prefer probability divergence metrics, e.g. Kullback-Leibler divergence, total-variation distance, χ^2 divergence

Mirror descent: adjust gradient updates to fit problem geometry

— Nemirovski & Yudin, '1983

Mirror descent (MD)

Replace the quadratic proximity $\|x - x^t\|_2^2$ with distance-like metric D_φ



$$x^{t+1} = \arg \min_{x \in \mathcal{C}} \left\{ f(x^t) + \langle \nabla f(x^t), x - x^t \rangle + \frac{1}{\eta_t} \underbrace{D_\varphi(x, x^t)}_{\text{Bregman divergence}} \right\}$$

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

Mirror descent (MD)

or more generally,

$$\mathbf{x}^{t+1} = \arg \min_{\mathbf{x} \in \mathcal{C}} \left\{ \cancel{f(\mathbf{x}^t)} + \langle \mathbf{g}^t, \mathbf{x} - \mathbf{x}^t \rangle + \frac{1}{\eta_t} D_{\varphi}(\mathbf{x}, \mathbf{x}^t) \right\} \quad (5.1)$$

with $\mathbf{g}^t \in \partial f(\mathbf{x}^t)$

- monitor local geometry via appropriate Bregman divergence metrics
 - generalization of squared Euclidean distance
 - e.g. squared Mahalanobis distance, KL divergence

Principles in choosing Bregman divergence

- fits the local curvature of $f(\cdot)$
- fits the geometry of the constraint set \mathcal{C}
- makes sure the Bregman projection (defined later) is inexpensive

Bregman divergence

Bregman divergence

Let $\varphi : \mathcal{C} \mapsto \mathbb{R}$ be strictly convex and differentiable on \mathcal{C} , then

$$D_{\varphi}(\mathbf{x}, \mathbf{z}) := \varphi(\mathbf{x}) - \varphi(\mathbf{z}) - \langle \nabla \varphi(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle$$

- shares a few similarities with squared Euclidean distance
- **a locally quadratic measure**: think of it as

$$D_{\varphi}(\mathbf{x}, \mathbf{z}) = (\mathbf{x} - \mathbf{z})^{\top} \nabla^2 \varphi(\boldsymbol{\xi})(\mathbf{x} - \mathbf{z})$$

for some $\boldsymbol{\xi}$ depending on \mathbf{x} and \mathbf{z}

Example: squared Mahalanobis distance

Let $D_\varphi(\mathbf{x}, \mathbf{z}) = \frac{1}{2}(\mathbf{x} - \mathbf{z})^\top \mathbf{Q}(\mathbf{x} - \mathbf{z})$ for $\mathbf{Q} \succ \mathbf{0}$, which is generated by

$$\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x}$$

Proof:

$$\begin{aligned} D_\varphi(\mathbf{x}, \mathbf{z}) &= \varphi(\mathbf{x}) - \varphi(\mathbf{z}) - \langle \nabla \varphi(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle \\ &= \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} - \frac{1}{2}\mathbf{z}^\top \mathbf{Q}\mathbf{z} - \mathbf{z}^\top \mathbf{Q}(\mathbf{x} - \mathbf{z}) \\ &= \frac{1}{2}(\mathbf{x} - \mathbf{z})^\top \mathbf{Q}(\mathbf{x} - \mathbf{z}) \end{aligned}$$



Example: squared Mahalanobis distance

When $D_\varphi(\mathbf{x}, \mathbf{z}) = \frac{1}{2}(\mathbf{x} - \mathbf{z})^\top \mathbf{Q}(\mathbf{x} - \mathbf{z})$, $\mathcal{C} = \mathbb{R}^n$, and f differentiable, MD has a closed-form expression

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \eta_t \mathbf{Q}^{-1} \nabla f(\mathbf{x}^t)$$

In general,

$$\begin{aligned} \mathbf{x}^{t+1} &= \arg \min_{\mathbf{x} \in \mathcal{C}} \left\{ \eta_t \langle \mathbf{g}^t, \mathbf{x} \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{x}^t)^\top \mathbf{Q}(\mathbf{x} - \mathbf{x}^t) \right\} \\ &= \arg \min_{\mathbf{x} \in \mathcal{C}} \left\{ \frac{1}{2} \mathbf{x}^\top \mathbf{Q} \mathbf{x} - \left\langle \mathbf{Q}(\mathbf{x}^t - \eta_t \mathbf{Q}^{-1} \mathbf{g}^t), \mathbf{x} \right\rangle + \cancel{\frac{1}{2} \mathbf{x}^{t\top} \mathbf{Q} \mathbf{x}^t} \right\} \\ &= \arg \min_{\mathbf{x} \in \mathcal{C}} \underbrace{\left\{ \frac{1}{2}(\mathbf{x} - (\mathbf{x}^t - \eta_t \mathbf{Q}^{-1} \mathbf{g}^t))^\top \mathbf{Q}(\mathbf{x} - (\mathbf{x}^t - \eta_t \mathbf{Q}^{-1} \mathbf{g}^t)) \right\}}_{\text{projection of } \mathbf{x}^t - \eta_t \mathbf{Q}^{-1} \mathbf{g}^t \text{ based on the weighted } \ell_2 \text{ distance } \|\mathbf{z}\|_Q^2 := \mathbf{z}^\top \mathbf{Q} \mathbf{z}} \end{aligned}$$

Example: KL divergence

Let $D_\varphi(\mathbf{x}, \mathbf{z}) = \text{KL}(\mathbf{x} \parallel \mathbf{z}) := \sum_i x_i \log \frac{x_i}{z_i}$, which is generated by

$$\varphi(\mathbf{x}) = \sum_i x_i \log x_i \quad (\text{negative entropy})$$

if $\mathcal{C} = \Delta := \{\mathbf{x} \in \mathbb{R}_+^n \mid \sum_i x_i = 1\}$ is the probability simplex

Proof:

$$\begin{aligned} D_\varphi(\mathbf{x}, \mathbf{z}) &= \varphi(\mathbf{x}) - \varphi(\mathbf{z}) - \langle \nabla \varphi(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle \\ &= \sum_i x_i \log x_i - \sum_i z_i \log z_i - \sum_i (\log z_i + 1)(x_i - z_i) \\ &= -\underbrace{\sum_i x_i}_{=1} + \underbrace{\sum_i z_i}_{=1} + \sum_i x_i \log \frac{x_i}{z_i} = \text{KL}(\mathbf{x} \parallel \mathbf{z}) \end{aligned}$$



Example: KL divergence

When $D_\varphi(\mathbf{x}, \mathbf{z}) = \text{KL}(\mathbf{x} \parallel \mathbf{z})$, $\mathcal{C} = \Delta$, and f differentiable, MD has closed-form (homework)

$$x_i^{t+1} = \frac{x_i^t \exp(-\eta_t [\nabla f(\mathbf{x}^t)]_i)}{\sum_{j=1}^n x_j^t \exp(-\eta_t [\nabla f(\mathbf{x}^t)]_j)}, \quad 1 \leq i \leq n$$

- often called **exponentiated gradient descent** or **entropic descent**

Example: generalized KL divergence

If $\mathcal{C} = \mathbb{R}_+^n$ (positive orthant), then the negative entropy $\varphi(\mathbf{x}) = \sum_i x_i \log x_i$ generates

$$D_\varphi(\mathbf{x}, \mathbf{z}) = \text{KL}(\mathbf{x} \parallel \mathbf{z}) := \sum_i x_i \log \frac{x_i}{z_i} - x_i + z_i$$

Example: von Neumann divergence

If $\mathcal{C} = \mathbb{S}_+^n$ (positive-definite cone), then the generalized negative entropy of eigenvalues

$$\varphi(\mathbf{X}) = \sum_i \lambda_i(\mathbf{X}) \log \lambda_i(\mathbf{X}) - \lambda_i(\mathbf{X}) =: \text{Tr}(\mathbf{X} \log \mathbf{X} - \mathbf{X})$$

generates the von Neumann divergence (commonly used in quantum mechanics)

$$\begin{aligned} D_\varphi(\mathbf{X}, \mathbf{Z}) &= \sum_i \lambda_i(\mathbf{X}) \log \frac{\lambda_i(\mathbf{X})}{\lambda_i(\mathbf{Z})} - \lambda_i(\mathbf{X}) + \lambda_i(\mathbf{Z}) \\ &=: \text{Tr}(\mathbf{X}(\log \mathbf{X} - \log \mathbf{Z}) - \mathbf{X} + \mathbf{Z}) \end{aligned}$$

Common families of Bregman divergence

Function Name	$\varphi(x)$	$\text{dom } \varphi$	$D_\varphi(x; y)$
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	$-x^p \quad (0 < p < 1)$	$[0, +\infty)$	$-x^p + pxy^{p-1} - (p-1)y^p$
ℓ_p norm	$ x ^p \quad (1 < p < \infty)$	$(-\infty, +\infty)$	$ x ^p - px \operatorname{sgn} y y ^{p-1} + (p-1) y ^p$
Exponential	$\exp x$	$(-\infty, +\infty)$	$\exp x - (x - y + 1) \exp y$
Inverse	$1/x$	$(0, +\infty)$	$1/x + x/y^2 - 2/y$

taken from I. Dhillon & J. Tropp, 2007

Basic properties of Bregman divergence

Let $\varphi : \mathcal{C} \mapsto \mathbb{R}$ be μ -strongly convex and differentiable on \mathcal{C}

- **non-negativity:** $D_\varphi(\mathbf{x}, \mathbf{z}) \geq 0$, and $D_\varphi(\mathbf{x}, \mathbf{z}) = 0$ iff $\mathbf{x} = \mathbf{z}$
 - in fact, $D_\varphi(\mathbf{x}, \mathbf{z}) \geq \frac{\mu}{2} \|\mathbf{x} - \mathbf{z}\|_2^2$ (by strong convexity of φ)
- **convexity:** $D_\varphi(\mathbf{x}, \mathbf{z})$ is $\underbrace{\text{convex in } \mathbf{x}}$, but not necessarily convex in \mathbf{z}
by defn, since φ is cvx
- **lack of symmetry:** in general, $D_\varphi(\mathbf{x}, \mathbf{z}) \neq D_\varphi(\mathbf{z}, \mathbf{x})$

Basic properties of Bregman divergence

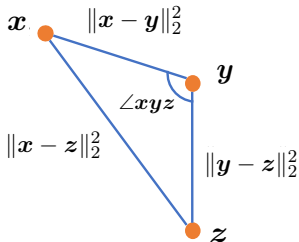
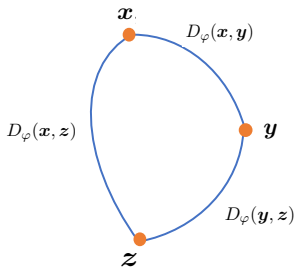
Let $\varphi : \mathcal{C} \mapsto \mathbb{R}$ be μ -strongly convex and differentiable on \mathcal{C}

- **linearity:** for φ_1, φ_2 strictly convex and $\lambda \geq 0$,

$$D_{\varphi_1 + \lambda \varphi_2}(\mathbf{x}, \mathbf{z}) = D_{\varphi_1}(\mathbf{x}, \mathbf{z}) + \lambda D_{\varphi_2}(\mathbf{x}, \mathbf{z})$$

- **unaffected by linear terms:** let $\varphi_2(\mathbf{x}) = \varphi_1(\mathbf{x}) + \mathbf{a}^\top \mathbf{x} + b$, then $D_{\varphi_2} = D_{\varphi_1}$
- **gradient:** $\nabla_{\mathbf{x}} D_{\varphi}(\mathbf{x}, \mathbf{z}) = \nabla \varphi(\mathbf{x}) - \nabla \varphi(\mathbf{z})$

Three-point lemma



Fact 5.1

For every three points x, y, z ,

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

- for Euclidean case with $\varphi(x) = \|x\|_2^2$, this is the **law of cosine**

$$\|x - z\|_2^2 = \|x - y\|_2^2 + \|y - z\|_2^2 - 2 \underbrace{\langle z - y, x - y \rangle}_{\|z - y\|_2 \|x - y\|_2 \cos \angle zyx}$$

Proof of the three-point lemma

$$\begin{aligned} & D_\varphi(\mathbf{x}, \mathbf{y}) + D_\varphi(\mathbf{y}, \mathbf{z}) - D_\varphi(\mathbf{x}, \mathbf{z}) \\ &= \varphi(\mathbf{x}) - \varphi(\mathbf{y}) - \langle \nabla \varphi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \varphi(\mathbf{y}) - \varphi(\mathbf{z}) - \langle \nabla \varphi(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle \\ &\quad - \{ \varphi(\mathbf{x}) - \varphi(\mathbf{z}) - \langle \nabla \varphi(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle \} \\ &= -\langle \nabla \varphi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle - \langle \nabla \varphi(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle + \langle \nabla \varphi(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle \\ &= \langle \nabla \varphi(\mathbf{z}) - \nabla \varphi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \end{aligned}$$

(Optional) connection with exponential families

Exponential family: a family of distributions with probability density (parametrized by $\boldsymbol{\theta}$)

$$p_{\varphi}(\mathbf{x} \mid \boldsymbol{\theta}) = \exp \{ \langle \mathbf{x}, \boldsymbol{\theta} \rangle - \varphi(\boldsymbol{\theta}) - h(\mathbf{x}) \}$$

for some cumulant function φ and some function h

- example (spherical Gaussian)

$$p_{\varphi}(\mathbf{x} \mid \boldsymbol{\theta}) \propto \exp \left\{ -\frac{\|\mathbf{x} - \boldsymbol{\theta}\|_2^2}{2} \right\} = \exp \left\{ \langle \mathbf{x}, \boldsymbol{\theta} \rangle - \underbrace{\frac{1}{2}\|\boldsymbol{\theta}\|_2^2}_{=:\varphi(\boldsymbol{\theta})} - \frac{\|\mathbf{x}\|_2^2}{2} \right\}$$

(Optional) connection with exponential families

For exponential families, under mild conditions, \exists function g_{φ^*} s.t.

$$p_{\varphi}(\mathbf{x} \mid \boldsymbol{\theta}) = \exp \{ -D_{\varphi^*}(\mathbf{x}, \boldsymbol{\mu}(\boldsymbol{\theta})) \} g_{\varphi^*}(\mathbf{x}) \quad (5.2)$$

where $\varphi^*(\boldsymbol{\theta}) := \sup_{\mathbf{x}} \{ \langle \mathbf{x}, \boldsymbol{\theta} \rangle - \varphi(\mathbf{x}) \}$ is the **Fenchel conjugate** of φ ,
and $\boldsymbol{\mu}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{x}]$

- \exists unique Bregman divergence associated with every member of exponential family

$$p_{\varphi}(\mathbf{x} \mid \boldsymbol{\theta}) \propto \exp \left\{ - \underbrace{\frac{\|\mathbf{x} - \boldsymbol{\mu}\|_2^2}{2}}_{D_{\varphi^*}(\mathbf{x}, \boldsymbol{\mu})} \right\}$$

(Optional) connection with exponential families

For exponential families, under mild conditions, \exists function g_{φ^*} s.t.

$$p_{\varphi}(\mathbf{x} \mid \boldsymbol{\theta}) = \exp \{ -D_{\varphi^*}(\mathbf{x}, \boldsymbol{\mu}(\boldsymbol{\theta})) \} g_{\varphi^*}(\mathbf{x}) \quad (5.2)$$

where $\varphi^*(\boldsymbol{\theta}) := \sup_{\mathbf{x}} \{ \langle \mathbf{x}, \boldsymbol{\theta} \rangle - \varphi(\mathbf{x}) \}$ is the **Fenchel conjugate** of φ , and $\boldsymbol{\mu}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{x}]$

- example (spherical Gaussian): since $\varphi^*(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, we have $D_{\varphi^*}(\mathbf{x}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{x} - \boldsymbol{\mu}\|_2^2$, which implies

$$p_{\varphi}(\mathbf{x} \mid \boldsymbol{\theta}) \propto \exp \left\{ - \underbrace{\frac{\|\mathbf{x} - \boldsymbol{\mu}\|_2^2}{2}}_{D_{\varphi^*}(\mathbf{x}, \boldsymbol{\mu})} \right\}$$

Proof of (5.2)

$$\begin{aligned} p_{\varphi}(\mathbf{x} \mid \boldsymbol{\theta}) &= \exp\{\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \varphi(\boldsymbol{\theta}) - h(\mathbf{x})\} \\ &\stackrel{(i)}{=} \exp\{\varphi^*(\boldsymbol{\mu}) + \langle \mathbf{x} - \boldsymbol{\mu}, \nabla \varphi^*(\boldsymbol{\mu}) \rangle - h(\mathbf{x})\} \\ &= \exp\{-\varphi^*(\mathbf{x}) + \varphi^*(\boldsymbol{\mu}) + \langle \mathbf{x} - \boldsymbol{\mu}, \nabla \varphi^*(\boldsymbol{\mu}) \rangle\} \exp\{\varphi^*(\mathbf{x}) - h(\mathbf{x})\} \\ &= \exp(-D_{\varphi^*}(\mathbf{x}, \boldsymbol{\mu})) \underbrace{\exp\{\varphi^*(\mathbf{x}) - h(\mathbf{x})\}}_{=: g_{\varphi^*}(\mathbf{x})} \end{aligned}$$

Here, (i) follows since (a) in exponential families, one has $\boldsymbol{\mu} = \nabla \varphi(\boldsymbol{\theta})$ and $\nabla \varphi^*(\boldsymbol{\mu}) = \boldsymbol{\theta}$, and (b) $\langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle = \varphi(\boldsymbol{\theta}) + \varphi^*(\boldsymbol{\mu})$ (homework)

Bregman projection

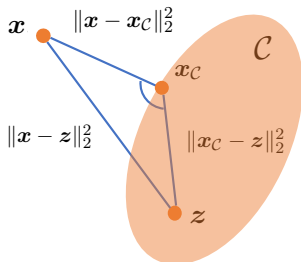
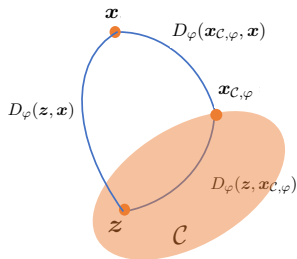
Given a point \boldsymbol{x} , define

$$\mathcal{P}_{\mathcal{C},\varphi}(\boldsymbol{x}) := \arg \min_{\boldsymbol{z} \in \mathcal{C}} D_{\varphi}(\boldsymbol{z}, \boldsymbol{x})$$

as the Bregman projection of \boldsymbol{x} onto \mathcal{C}

- as we shall see, MD is useful when Bregman projection requires little computational effort

Generalized Pythagorean Theorem



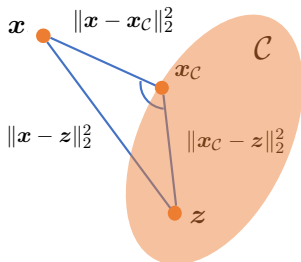
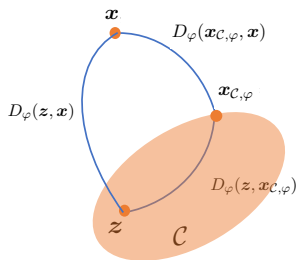
Fact 5.2

If $x_{C,\varphi} = \mathcal{P}_{C,\varphi}(x)$, then

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

- in the squared Euclidean case, it means the angle $\angle z x_C x$ is *obtuse*

Generalized Pythagorean Theorem



Fact 5.2

If $x_{C,\varphi} = \mathcal{P}_{C,\varphi}(x)$, then

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

- if C is an **affine plane**, then

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

Proof of Fact 5.2

Let

$$\mathbf{g} = \nabla_z D_\varphi(\mathbf{z}, \mathbf{x}) \Big|_{\mathbf{z}=\mathbf{x}_{\mathcal{C},\varphi}} = \nabla\varphi(\mathbf{x}_{\mathcal{C},\varphi}) - \nabla\varphi(\mathbf{x})$$

Since $\mathbf{x}_{\mathcal{C},\varphi} = \arg \min_{\mathbf{z} \in \mathcal{C}} D_\varphi(\mathbf{z}, \mathbf{x})$, the optimality condition for constrained convex optimization gives (see Bertsekas '16)

$$\langle \mathbf{g}, \mathbf{z} - \mathbf{x}_{\mathcal{C},\varphi} \rangle \geq 0 \quad \forall \mathbf{z} \in \mathcal{C}$$

Therefore, for all $\mathbf{z} \in \mathcal{C}$,

$$\begin{aligned} 0 &\geq \langle \mathbf{g}, \mathbf{x}_{\mathcal{C},\varphi} - \mathbf{z} \rangle = \langle \nabla\varphi(\mathbf{x}) - \nabla\varphi(\mathbf{x}_{\mathcal{C},\varphi}), \mathbf{z} - \mathbf{x}_{\mathcal{C},\varphi} \rangle \\ &= D_\varphi(\mathbf{z}, \mathbf{x}_{\mathcal{C},\varphi}) + D_\varphi(\mathbf{x}_{\mathcal{C},\varphi}, \mathbf{x}) - D_\varphi(\mathbf{z}, \mathbf{x}) \end{aligned}$$

as claimed, where the last line comes from Fact 5.1

Alternative forms of mirror descent

An alternative form of MD

Using the Bregman divergence, one can also describe MD as

$$\nabla\varphi(\mathbf{y}^{t+1}) = \nabla\varphi(\mathbf{x}^t) - \eta_t \mathbf{g}^t \quad \text{with } \mathbf{g}^t \in \partial f(\mathbf{x}^t) \quad (5.3a)$$

$$\mathbf{x}^{t+1} \in \mathcal{P}_{\mathcal{C},\varphi}(\mathbf{y}^{t+1}) = \arg \min_{\mathbf{z} \in \mathcal{C}} D_{\varphi}(\mathbf{z}, \mathbf{y}^{t+1}) \quad (5.3b)$$

- performs gradient descent in certain “dual” space

An alternative form of MD

The equivalence can be seen by looking at the optimality conditions

- the optimality condition of (5.3b) gives

$$\begin{aligned} \mathbf{0} &\in \nabla\varphi(\mathbf{x}^{t+1}) - \nabla\varphi(\mathbf{y}^{t+1}) + \underbrace{N_{\mathcal{C}}(\mathbf{x}^{t+1})}_{\text{normal cone}} \quad (\text{see Bertsekas '16}) \\ &= \varphi(\mathbf{x}^{t+1}) - \nabla\varphi(\mathbf{x}^t) + \eta_t \mathbf{g}^t + N_{\mathcal{C}}(\mathbf{x}^{t+1}) \end{aligned} \quad (5.3a)$$

- the optimality condition of (5.1) reads

$$\mathbf{0} \in \mathbf{g}^t + \frac{1}{\eta_t} \left\{ \nabla\varphi(\mathbf{x}^{t+1}) - \nabla\varphi(\mathbf{x}^t) \right\} + N_{\mathcal{C}}(\mathbf{x}^{t+1}) \quad (\text{see Bertsekas '16})$$

- these two conditions are clearly identical

Another form of MD

For simplicity, assume $\mathcal{C} = \mathbb{R}^n$, then another form is

$$\mathbf{x}^{t+1} = \nabla \varphi^* \left(\nabla \varphi(\mathbf{x}^t) - \eta \mathbf{g}^t \right) \quad (5.4)$$

where $\varphi^*(\mathbf{x}) := \sup_{\mathbf{z}} \{ \langle \mathbf{z}, \mathbf{x} \rangle - \varphi(\mathbf{z}) \}$ is the Fenchel-conjugate of φ

- this is the version originally proposed in Nemirovski & Yudin '1983

Another form of MD

When $\mathcal{C} = \mathbb{R}^n$, (5.3a)-(5.3b) simplifies to

$$\mathbf{x}^{t+1} = \mathbf{y}^{t+1} = (\nabla\varphi)^{-1}\left(\nabla\varphi(\mathbf{x}^t) - \eta\mathbf{g}^t\right)$$

It thus suffices to show

$$(\nabla\varphi)^{-1} = (\nabla\varphi)^* \tag{5.5}$$

Proof of Claim (5.5)

Suppose $\mathbf{y} = \nabla\varphi(\mathbf{x})$. From the conjugate subgradient theorem, this is equivalent to (homework)

$$\varphi(\mathbf{x}) + \varphi^*(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$$

Since $\varphi^{**} = \varphi$, we further have

$$\varphi^*(\mathbf{y}) + \varphi^{**}(\mathbf{x}) = \langle \mathbf{x}, \mathbf{y} \rangle,$$

which combined with the conjugate subgradient theorem yields $\mathbf{x} = \nabla\varphi^*(\mathbf{y})$. This means

$$\mathbf{x} = \nabla\varphi^*(\mathbf{y}) = \nabla\varphi^*(\nabla\varphi(\mathbf{x}))$$

and hence $\nabla\varphi^* = (\nabla\varphi)^{-1}$

Convergence analysis

Convex and Lipschitz problems

$$\begin{array}{ll}\text{minimize}_x & f(x) \\ \text{subject to} & x \in \mathcal{C}\end{array}$$

- f is convex and Lipschitz continuous
 - φ is ρ -strongly convex w.r.t. a certain norm $\|\cdot\|$
 - $\|\mathbf{g}\|_* \leq L_f$ for any subgradient $\mathbf{g} \in \partial f(x)$ at any point x , where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$

Convergence analysis

Theorem 5.3

Suppose f is convex and Lipschitz continuous (in the sense that $\|g\|_* \leq L_f$ for any subgradient g of f) on \mathcal{C} . Suppose φ is ρ -strongly convex w.r.t. $\|\cdot\|$. Then

$$f^{\text{best},t} - f^{\text{opt}} \leq \frac{\sup_{\mathbf{x} \in \mathcal{C}} D_\varphi(\mathbf{x}, \mathbf{x}^0) + \frac{L_f^2}{2\rho} \sum_{k=0}^t \eta_k^2}{\sum_{k=0}^t \eta_k}$$

- If $\eta_t = \frac{\sqrt{2\rho R}}{L_f} \frac{1}{\sqrt{t}}$ with $R := \sup_{\mathbf{x} \in \mathcal{C}} D_\varphi(\mathbf{x}, \mathbf{x}^0)$, then

$$f^{\text{best},t} - f^{\text{opt}} \leq O\left(\frac{L_f \sqrt{R}}{\sqrt{\rho}} \frac{\log t}{\sqrt{t}}\right)$$

- one can further remove the $\log t$ factor

Example: optimization over probability simplex

Suppose $\mathcal{C} = \Delta$ is the probability simplex, and pick $\mathbf{x}^0 = n^{-1}\mathbf{1}$

(1) set $\varphi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, which is 1-strongly convex w.r.t. $\|\cdot\|_2$. Then

$$\sup_{\mathbf{x} \in \Delta} D_{\varphi}(\mathbf{x}, \mathbf{x}^0) = \sup_{\mathbf{x} \in \Delta} \frac{1}{2}\|\mathbf{x} - n^{-1}\mathbf{1}\|_2^2 = \sup_{\mathbf{x} \in \Delta} \frac{1}{2}\left(\|\mathbf{x}\|_2^2 - \frac{1}{n}\right) \leq \frac{1}{2}$$

Then Theorem 5.3 says

$$f^{\text{best},t} - f^{\text{opt}} \leq O\left(L_{f,2} \frac{\log t}{\sqrt{t}}\right)$$

if any subgradient \mathbf{g} obeys $\|\mathbf{g}\|_2 \leq L_{f,2}$

Example: optimization over probability simplex

Suppose $\mathcal{C} = \Delta$ is the probability simplex, and pick $\mathbf{x}^0 = n^{-1}\mathbf{1}$

(2) set $\phi(\mathbf{x}) = -\sum_{i=1}^n x_i \log x_i$, which is 1-strongly convex w.r.t. $\|\cdot\|_1$. Then

$$\begin{aligned}\sup_{\mathbf{x} \in \Delta} D_{\phi}(\mathbf{x}, \mathbf{x}^0) &= \sup_{\mathbf{x} \in \Delta} \text{KL}(\mathbf{x} \parallel \mathbf{x}^0) = \sup_{\mathbf{x} \in \Delta} \sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n x_i \log \frac{1}{n} \\ &= \log n + \sup_{\mathbf{x} \in \Delta} \sum_{i=1}^n x_i \log x_i \leq \log n\end{aligned}$$

Then Theorem 5.3 says

$$f^{\text{best},t} - f^{\text{opt}} \leq O\left(L_{f,\infty} \sqrt{\log n} \frac{\log t}{\sqrt{t}}\right)$$

if any subgradient \mathbf{g} obeys $\|\mathbf{g}\|_{\infty} \leq L_{f,\infty}$

Example: optimization over probability simplex

Comparing these two choices and ignoring log terms, we have

$$\text{Euclidean: } O\left(\frac{L_{f,2}}{\sqrt{t}}\right) \quad \text{vs.} \quad \text{KL: } O\left(\frac{L_{f,\infty}}{\sqrt{t}}\right)$$

Since $\|\mathbf{g}\|_\infty \leq \|\mathbf{g}\|_2 \leq \sqrt{n}\|\mathbf{g}\|_\infty$, one has

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

and hence the KL version often yields much better performance

Numerical example: robust regression

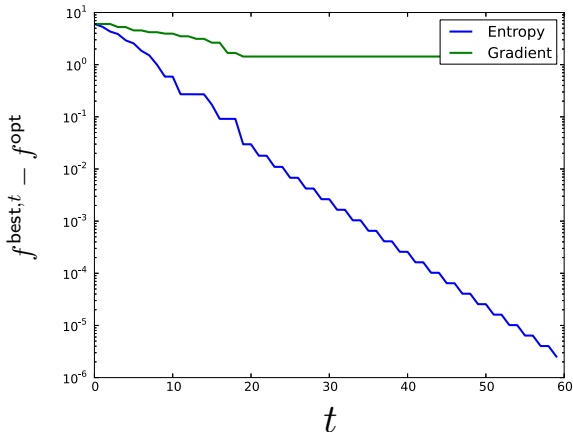
taken from Stanford EE364B

$$\begin{aligned} \text{minimize}_{\mathbf{x}} \quad & f(\mathbf{x}) = \sum_{i=1}^m |\mathbf{a}_i^\top \mathbf{x} - b_i| \\ \text{subject to} \quad & \mathbf{x} \in \Delta = \{\mathbf{x} \in \mathbb{R}_+^n \mid \mathbf{1}^\top \mathbf{x} = 1\} \end{aligned}$$

with $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n \times n})$ and $b_i = \frac{a_{i,1} + a_{i,2}}{2} + \mathcal{N}(0, 10^{-2})$, $m = 20$,
 $n = 3000$

Numerical example: robust regression

taken from Stanford EE364B



Fundamental inequality for mirror descent

Lemma 5.4

$$\eta_t \left(f(\mathbf{x}^t) - f^{\text{opt}} \right) \leq D_\varphi(\mathbf{x}^*, \mathbf{x}^t) - D_\varphi(\mathbf{x}^*, \mathbf{x}^{t+1}) + \frac{\eta_t^2 L_f^2}{2\rho}$$

- $D_\varphi(\mathbf{x}^*, \mathbf{x}^t) - D_\varphi(\mathbf{x}^*, \mathbf{x}^{t+1})$ motivates us to form a telescopic sum later

Proof of Lemma 5.4

$$\begin{aligned} f(\mathbf{x}^t) - f(\mathbf{x}^*) &\leq \langle \mathbf{g}^t, \mathbf{x}^t - \mathbf{x}^* \rangle && \text{(property of subgradient)} \\ &= \frac{1}{\eta_t} \langle \nabla \varphi(\mathbf{x}^t) - \nabla \varphi(\mathbf{y}^{t+1}), \mathbf{x}^t - \mathbf{x}^* \rangle && \text{(MD update rule)} \\ &= \frac{1}{\eta_t} \{ D_\varphi(\mathbf{x}^*, \mathbf{x}^t) + D_\varphi(\mathbf{x}^t, \mathbf{y}^{t+1}) - D_\varphi(\mathbf{x}^*, \mathbf{y}^{t+1}) \} && \text{(three point lemma)} \\ &\leq \frac{1}{\eta_t} \{ D_\varphi(\mathbf{x}^*, \mathbf{x}^t) + D_\varphi(\mathbf{x}^t, \mathbf{y}^{t+1}) - D_\varphi(\mathbf{x}^*, \mathbf{x}^{t+1}) - D_\varphi(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}) \} \\ &&& \text{(Pythagorean)} \\ &= \frac{1}{\eta_t} \{ D_\varphi(\mathbf{x}^*, \mathbf{x}^t) - D_\varphi(\mathbf{x}^*, \mathbf{x}^{t+1}) \} + \frac{1}{\eta_t} \{ D_\varphi(\mathbf{x}^t, \mathbf{y}^{t+1}) - D_\varphi(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}) \} \end{aligned}$$

so we need to first bound the 2nd term of the last line

Proof of Lemma 5.4 (cont.)

We claim that

$$D_{\varphi}(\mathbf{x}^t, \mathbf{y}^{t+1}) - D_{\varphi}(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}) \leq \frac{(\eta_t L_f)^2}{2\rho} \quad (5.6)$$

This gives

$$\eta_t (f(\mathbf{x}^t) - f(\mathbf{x}^*)) \leq \{D_{\varphi}(\mathbf{x}^*, \mathbf{x}^t) - D_{\varphi}(\mathbf{x}^*, \mathbf{x}^{t+1})\} + \frac{(\eta_t L_f)^2}{2\rho}$$

as claimed

Proof of Lemma 5.4 (cont.)

Finally, we justify (5.6):

$$\begin{aligned} & D_\varphi(\mathbf{x}^t, \mathbf{y}^{t+1}) - D_\varphi(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}) \\ &= \varphi(\mathbf{x}^t) - \varphi(\mathbf{x}^{t+1}) - \langle \nabla \varphi(\mathbf{y}^{t+1}), \mathbf{x}^t - \mathbf{x}^{t+1} \rangle \\ &\leq \langle \nabla \varphi(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}^{t+1} \rangle - \frac{\rho}{2} \|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 - \langle \nabla \varphi(\mathbf{y}^{t+1}), \mathbf{x}^t - \mathbf{x}^{t+1} \rangle \\ &\hspace{25em} \text{(strong convexity of } \varphi) \\ &= \langle \nabla \varphi(\mathbf{x}^t) - \nabla \varphi(\mathbf{y}^{t+1}), \mathbf{x}^t - \mathbf{x}^{t+1} \rangle - \frac{\rho}{2} \|\mathbf{x}^t - \mathbf{x}^{t+1}\|_2^2 \\ &= \eta_t \langle \mathbf{g}^t, \mathbf{x}^t - \mathbf{x}^{t+1} \rangle - \frac{\rho}{2} \|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 \hspace{2em} \text{(MD update rule)} \\ &\leq \eta_t L_f \|\mathbf{x}^t - \mathbf{x}^{t+1}\| - \frac{\rho}{2} \|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 \hspace{2em} \text{(Cauchy-Schwarz)} \\ &\leq \frac{(\eta_t L_f)^2}{2\rho} \hspace{10em} \text{(optimize quadratic function in } \|\mathbf{x}^t - \mathbf{x}^{t+1}\|) \end{aligned}$$

Proof of Theorem 5.3

From Lemma 5.4, one has

$$\eta_k \left(f(\mathbf{x}^k) - f^{\text{opt}} \right) \leq D_\varphi(\mathbf{x}^*, \mathbf{x}^k) - D_\varphi(\mathbf{x}^*, \mathbf{x}^{k+1}) + \frac{\eta_k^2 L_f^2}{2\rho}$$

Taking this inequality for $k = 0, \dots, t$ and summing them up give

$$\begin{aligned} \sum_{k=0}^t \eta_k \left(f(\mathbf{x}^k) - f^{\text{opt}} \right) &\leq D_\varphi(\mathbf{x}^*, \mathbf{x}^0) - D_\varphi(\mathbf{x}^*, \mathbf{x}^{t+1}) + \frac{L_f^2 \sum_{k=0}^t \eta_k^2}{2\rho} \\ &\leq \sup_{\mathbf{x} \in \mathcal{C}} D_\varphi(\mathbf{x}, \mathbf{x}^0) + \frac{L_f^2 \sum_{k=0}^t \eta_k^2}{2\rho} \end{aligned}$$

This together with $f^{\text{best},t} - f^{\text{opt}} \leq \frac{\sum_{k=0}^t \eta_k (f(\mathbf{x}^k) - f^{\text{opt}})}{\sum_{k=0}^t \eta_k}$ concludes the proof

Reference

- [1] "*Problem complexity and method efficiency in optimization*," A. Nemirovski, D. Yudin, Wiley, 1983.
- [2] "*Mirror descent and nonlinear projected subgradient methods for convex optimization*," A. Beck, M. Teboulle, Operations Research Letters, 31(3), 2003.
- [3] "*Convex optimization: algorithms and complexity*," S. Bubeck, Foundations and trends in machine learning, 2015.
- [4] "*First-order methods in optimization*," A. Beck, Vol. 25, SIAM, 2017.
- [5] "*Mathematical optimization, MATH301 lecture notes*," E. Candes, Stanford.
- [6] "*Convex optimization, EE364B lecture notes*," S. Boyd, Stanford.

Reference

- [7] "*Matrix nearness problems with Bregman divergences*," I. Dhillon, J. Tropp, SIAM Journal on Matrix Analysis and Applications, 29(4), 2007.
- [8] "*Nonlinear Programming (2nd Edition)*," D. Bertsekas, Athena Scientific, 2016.