Mirror descent

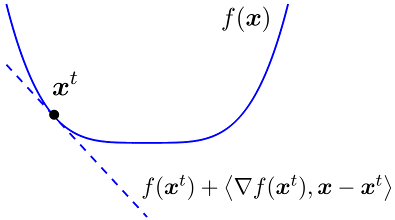
Yuxin Chen

Princeton University, Fall 2019

Outline

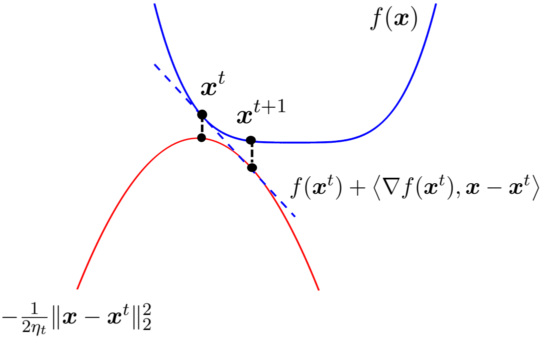
- Mirror descent
- Bregman divergence
- Alternative forms of mirror descent
- Convergence analysis

A proximal viewpoint of projected GD



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

A proximal viewpoint of projected GD



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

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

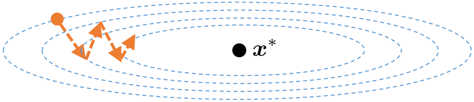
Inhomoneneous / non-Euclidean geometry

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

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

Example: quadratic minimization

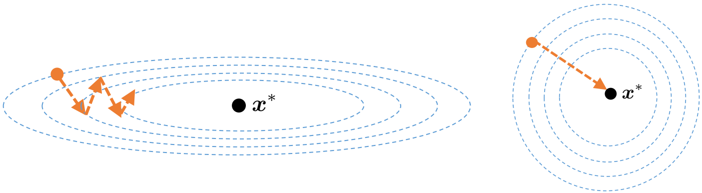


$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^*)^\top \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{x}^*)$$

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

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

Example: quadratic minimization



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

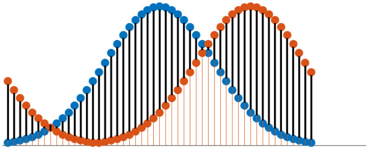
where $\boldsymbol{Q} \succ \boldsymbol{0}$ is a diagonal matrix with large $\kappa = \frac{\max_i Q_{i,i}}{\min_i Q_{i,i}} \gg 1$

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

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

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

Example: probability simplex



total-variation distance

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

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

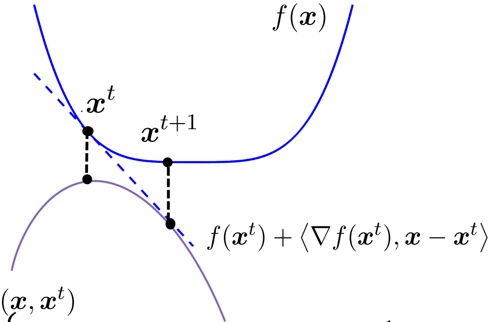
- Euclidean distance is in general not recommended for measuring the distance between probability vectors
- may prefer probability divergence metrics, e.g. Kullback-Leibler divergence, total-variation distance, χ^2 divergence

Mirror descent: adjust gradient updates to fit problem geometry

— Nemirovski & Yudin, '1983

Mirror descent (MD)

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



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

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

Mirror descent (MD)

or more generally,

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

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

- 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

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

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

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

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

Example: squared Mahalanobis distance

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

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

Proof:

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

$\square$

Example: squared Mahalanobis distance

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

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

In general,

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

Example: KL divergence

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

$$\varphi(\boldsymbol{x}) = \sum_i x_i \log x_i \qquad \text{(negative entropy)}$$

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

Proof:

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

$\square$

Example: KL divergence

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

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

- often called exponentiated gradient descent or entropic descent

Example: generalized KL divergence

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

$$D_\varphi(\boldsymbol{x}, \boldsymbol{z}) = \mathsf{KL}(\boldsymbol{x} \,\|\, \boldsymbol{z}) := \sum_i x_i \log \frac{x_i}{z_i} - x_i + z_i$$

Example: von Neumann divergence

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

$$\varphi(\boldsymbol{X}) = \sum_i \lambda_i(\boldsymbol{X}) \log \lambda_i(\boldsymbol{X}) - \lambda_i(\boldsymbol{X}) =: \mathrm{Tr}(\boldsymbol{X} \log \boldsymbol{X} - \boldsymbol{X})$$

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

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

Common families of Bregman divergence

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

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

Basic properties of Bregman divergence

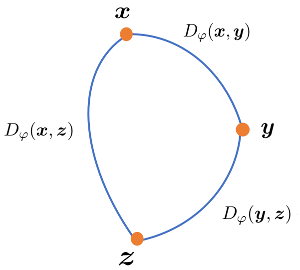
Let $\varphi : \mathcal{C} \mapsto \mathbb{R}$ be μ-strongly convex and differentiable on $\mathcal{C}$

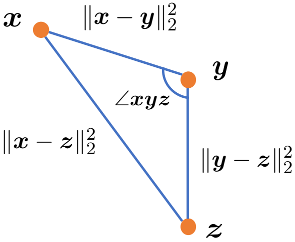
- **linearity:** for φ_1, φ_2 strictly convex and $\lambda \geq 0$,

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

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

Three-point lemma





Fact 5.1

For every three points $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$,

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

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

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

Proof of the three-point lemma

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

(Optional) connection with exponential families

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

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

for some cumulant function φ and some function h

- example (spherical Gaussian)

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

(Optional) connection with exponential families

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

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

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

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

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

(Optional) connection with exponential families

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

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

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

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

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

Proof of (5.2)

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

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

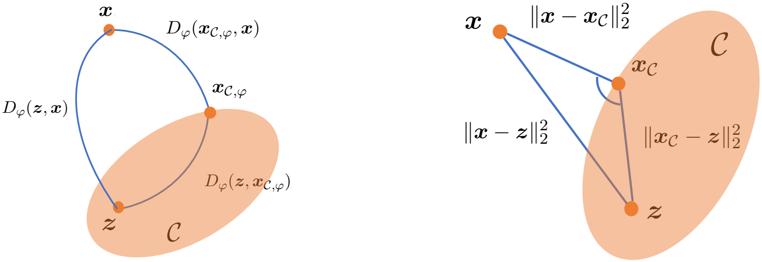
Given a point $\boldsymbol{x}$, define

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

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



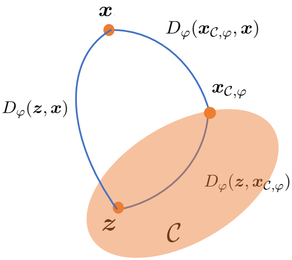
Fact 5.2

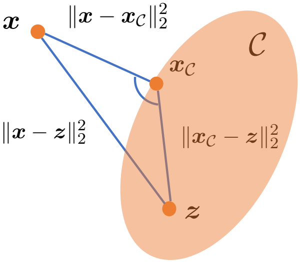
If $\boldsymbol{x}_{\mathcal{C},\varphi} = \mathcal{P}_{\mathcal{C},\varphi}(\boldsymbol{x})$, *then*

$$D_\varphi(\boldsymbol{z}, \boldsymbol{x}) \geq D_\varphi(\boldsymbol{z}, \boldsymbol{x}_{\mathcal{C},\varphi}) + D_\varphi(\boldsymbol{x}_{\mathcal{C},\varphi}, \boldsymbol{x}) \qquad \forall \boldsymbol{z} \in \mathcal{C}$$

- in the squared Euclidean case, it means the angle $\angle \boldsymbol{z}\boldsymbol{x}_{\mathcal{C},\varphi}\boldsymbol{x}$ is *obtuse*

Generalized Pythagorean Theorem





Fact 5.2

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

$$D_\varphi(\boldsymbol{z}, \boldsymbol{x}) \geq D_\varphi(\boldsymbol{z}, \boldsymbol{x}_{\mathcal{C},\varphi}) + D_\varphi(\boldsymbol{x}_{\mathcal{C},\varphi}, \boldsymbol{x}) \qquad \forall \boldsymbol{z} \in \mathcal{C}$$

- if $\mathcal{C}$ is an affine plane, then

$$D_\varphi(\boldsymbol{z}, \boldsymbol{x}) = D_\varphi(\boldsymbol{z}, \boldsymbol{x}_{\mathcal{C},\varphi}) + D_\varphi(\boldsymbol{x}_{\mathcal{C},\varphi}, \boldsymbol{x}) \qquad \forall \boldsymbol{z} \in \mathcal{C}$$

Proof of Fact 5.2

Let

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

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

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

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

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

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

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

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

- 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

$$\mathbf{0} \in \nabla\varphi(\boldsymbol{x}^{t+1}) - \nabla\varphi(\boldsymbol{y}^{t+1}) + \underbrace{N_{\mathcal{C}}(\boldsymbol{x}^{t+1})}_{\text{normal cone}} \quad \text{(see Bertsekas '16)}$$

$$= \varphi(\boldsymbol{x}^{t+1}) - \nabla\varphi(\boldsymbol{x}^{t}) + \eta_t \boldsymbol{g}^t + N_{\mathcal{C}}(\boldsymbol{x}^{t+1}) \qquad \text{(5.3a)}$$

- the optimality condition of (5.1) reads

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

- these two conditions are clearly identical

Another form of MD

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

$$\boldsymbol{x}^{t+1} = \nabla\varphi^*\Big(\nabla\varphi(\boldsymbol{x}^t) - \eta\boldsymbol{g}^t\Big) \tag{5.4}$$

where $\varphi^*(\boldsymbol{x}) := \sup_{\boldsymbol{z}}\{\langle \boldsymbol{z}, \boldsymbol{x}\rangle - \varphi(\boldsymbol{z})\}$ 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

$$\boldsymbol{x}^{t+1} = \boldsymbol{y}^{t+1} = (\nabla\varphi)^{-1}\Big(\nabla\varphi(\boldsymbol{x}^t) - \eta\boldsymbol{g}^t\Big)$$

It thus sufficies to show

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

Proof of Claim (5.5)

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

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

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

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

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

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

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

Convergence analysis

Convex and Lipschitz problems

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

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

Convergence analysis

Theorem 5.3

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

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

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

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

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

Example: optimization over probability simplex

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

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

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

Then Theorem 5.3 says

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

if any subgradient $\boldsymbol{g}$ obeys $\|\boldsymbol{g}\|_2 \le L_{f,2}$

Example: optimization over probability simplex

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

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

$$\sup_{\boldsymbol{x}\in\Delta} D_{\varphi}(\boldsymbol{x}, \boldsymbol{x}^0) = \sup_{\boldsymbol{x}\in\Delta} \mathsf{KL}(\boldsymbol{x} \,\|\, \boldsymbol{x}^0) = \sup_{\boldsymbol{x}\in\Delta} \sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n x_i \log \frac{1}{n}$$
$$= \log n + \sup_{\boldsymbol{x}\in\Delta} \sum\nolimits_{i=1}^n x_i \log x_i \le \log n$$

Then Theorem 5.3 says

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

if any subgradient $\boldsymbol{g}$ obeys $\|\boldsymbol{g}\|_\infty \le L_{f,\infty}$

Example: optimization over probability simplex

Comparing these two choices and ignoring log terms, we have

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

Since $\|\boldsymbol{g}\|_\infty \leq \|\boldsymbol{g}\|_2 \leq \sqrt{n}\|\boldsymbol{g}\|_\infty$, 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

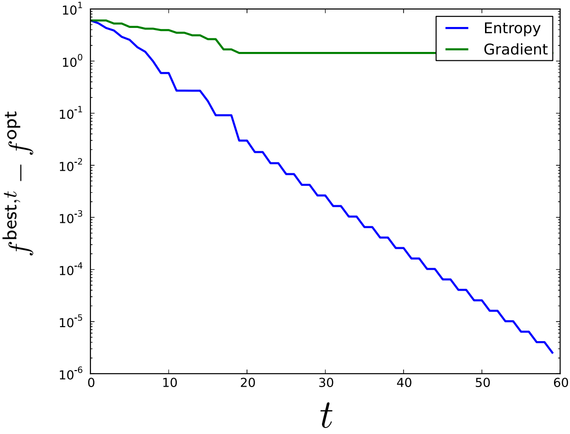
taken from Stanford EE364B

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

with $\boldsymbol{a}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{n\times n})$ 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

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

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

Proof of Lemma 5.4

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

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

Proof of Lemma 5.4 (cont.)

We claim that

$$D_\varphi\big(\boldsymbol{x}^t, \boldsymbol{y}^{t+1}\big) - D_\varphi\big(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}\big) \le \frac{(\eta_t L_f)^2}{2\rho} \tag{5.6}$$

This gives

$$\eta_t \left(f\big(\boldsymbol{x}^t\big) - f\big(\boldsymbol{x}^*\big)\right) \le \left\{D_\varphi\big(\boldsymbol{x}^*, \boldsymbol{x}^t\big) - D_\varphi\big(\boldsymbol{x}^*, \boldsymbol{x}^{t+1}\big)\right\} + \frac{(\eta_t L_f)^2}{2\rho}$$

as claimed

Proof of Lemma 5.4 (cont.)

Finally, we justify (5.6):

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

Proof of Theorem 5.3

From Lemma 5.4, one has

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

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

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

This together with $f^{\mathsf{best},t} - f^{\mathsf{opt}} \le \frac{\sum_{k=0}^{t} \eta_k \left(f(\boldsymbol{x}^k) - f^{\mathrm{opt}} \right)}{\sum_{k=0}^{t} \eta_k}$ 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.