Gauge optimization and duality

Junfeng Yang

Department of Mathematics
Nanjing University

Joint with Shiqian Ma, CUHK

September, 2015
Outline

Introduction

Duality
  Lagrange duality
  Fenchel duality
  Gauge duality
Outline

Introduction

Duality
  Lagrange duality
  Fenchel duality
  Gauge duality
Gauge function

Definition (gauge function)
A function $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ is called a gauge function if

1. $f$ is convex: for all $x, y \in \mathbb{R}^n$ and $\alpha \in (0, 1)$,
   \[ f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y); \]

2. $f$ is positively homogeneous: for all $x \in \mathbb{R}^n$ and $\lambda > 0$,
   \[ f(\lambda x) = \lambda f(x); \]

3. $f$ is nonnegative: $f(x) \geq 0$ for all $x$;

4. $f$ vanishes at the origin: $f(0) = 0$.

Norms and semi-norms are particular gauge functions.
The problem

We are interested in the following gauge optimization

$$\min_{x} \{ \kappa(x) \mid \rho(Ax - b) \leq \varepsilon \},$$

- $\kappa$ and $\rho$ are gauge functions, and $\rho^{-1}(0) = 0$;
- $A$, $b$ and $\varepsilon > 0$ are given data.
Sparsity

Original (n = 4096, number of nonzeros = 204)
Sparse solutions

**Problem:** find a sparse solution $x$

$$\min_{x} \{ \| x \|_0 \mid Ax \approx b \},$$

or, find $x$ such that the residue is sparse

$$\min_{x} \| Ax - b \|_0.$$

**Drawback:** Combinatorial optimization, difficult.
Sparse representation

- represent a signal \( y \) as a superposition of elementary signal atoms:

\[
y = \sum_i \phi_i x_i \quad \text{or} \quad y = \Phi x;
\]

- unique representation if the basis is orthonormal;
- overcomplete dictionaries have more flexibility, but representation is not unique;
- a "good" transform leads to a fast decay in coefficients.
Sparse representation

\[ y = Hx \]
\[ y = Dx \]
\[ y = [H \quad D]x \]

Haar  
DCT  
Haar/DCT dictionary
Sparse representation
Sparse representation

Identity

Fourier

Wavelet
Sparse representation

Identity

Fourier

Wavelet
Sparse representation

Identity

Fourier

Wavelet
Sparse representation

Identity

Fourier

Wavelet
Sparse representation

Identity

Fourier

Wavelet
Sparse representation

Identity  Fourier  Wavelet
The LASSO model (Tibshirani, 1996)

\[ \min_x \{ \|Ax - b\|_2 \mid \|x\|_1 \leq k \} . \]

The BP/BPDN model (Donoho et al., 1998; Candes et al., 2006)

\[ \min_x \{ \|x\|_1 \mid \|Ax - b\|_2 \leq \varepsilon \} . \]

Joint sparsity (Malioutov-Cetin-Wilsky 2003)

\[ \min_X \left\{ \sum_j \|X(:,j)\|_2 \mid \|AX - B\|_2 \leq \varepsilon \right\} . \]

The last two are gauge optimization.
Matrix completion

Seek a matrix $X \in \mathbb{R}^{m \times n}$ with some desired structure (e.g., low rank) that matches certain observations, possibly noisy.

$$\min_{X} \{ \kappa(X) \mid \|AX - b\| \leq \varepsilon \},$$

where $A$ is a linear mapping (e.g. observations of certain elements of $X$).

Setting $\kappa$ as the nuclear norm, i.e., sum of all singular values, promotes low rank solutions. Applications in recommender systems, e.g., Netflix, Amazon. (Recht et al., 2010)
Phase retrieval

In phase retrieval, one recovers the phase information of a signal, e.g., an image, from magnitude-only measurements (Candes et al. 2012, Waldspurger 2015). Application in X-ray crystallography, used to image the molecular structure of a crystal (Harrison, 1993).

- unknown signal $x \in \mathbb{C}^n$;
- measurements are given by

$$b_k = |\langle x, a_k \rangle|^2$$

for some vectors $a_k$ that encode the waveforms used to illuminate the signal, $k = 1, \ldots, m$. 
Phase Lift

These measurements can be understood as

\[ b_k = \langle xx^*, a_k a_k^* \rangle = \langle X, A_k \rangle \]

of the lifted signal \( X := xx^* \), where \( A_k := a_k a_k^* \) is the \( k \)-th lifted rank-1 measurement matrix.

The Phase Lift convex model

\[
\min_{X \in S^n} \left\{ \langle I, X \rangle + \iota_{\geq 0}(X) \mid \|AX - b\| \leq \varepsilon \right\}
\]

also falls into the class of gauge optimization problems.
Image reconstruction

In image reconstruction, one recovers an unknown image from a set of linear measurements

\[ b = Ax^* + \text{noise}. \]

The total variation model (Rudin-Osher-Fatemi, 1992)

\[
\min_x \{ \text{TV}(x) \mid \|Ax - b\| \leq \varepsilon \},
\]

where \( \text{TV}(x) := \sum_i \|D_i x\| \) and \( A \) is

- identity (denoising)
- blurring operator (deconvolution)
- partial convolution/wavelet (inpainting)
- partial Fourier (MRI)
In deconvolution/deblurring, $A$ is a convolution matrix (square and invertible). The *least squares estimation* (LSE) of $\tilde{x}$ is

$$x^{LSE} = \arg \min_x \|Ax - f\|$$

Left: $f$; Right: $x^{LSE}$. 
Ill-posed

In MRI, observed data satisfy $f = P \mathcal{F} \tilde{x} + \text{noise}$:

- $\mathcal{F}$: 2D Fourier matrix
- $P$: projection matrix

Left: $k$-space sampling (9.36%); Right: back projection.
RGB image recovery

Model: \( \min_x \{ \text{MTV}(x) \mid \|Kx - f\| \leq \varepsilon \} \).

SNR 6.25dB

SNR: 17.53dB, CPU 2.78s, It: 10,

Cross-channel blur, Gaussian noise, 256 \times 256.
Deconvolution with impulsive noise

Model: \( \min_x \{ \text{MTV}(x) \mid \|Kx - f\|_1 \leq \varepsilon \} \).

Corruption: 60%

SNR 11.74dB, CPU 18.06s, It 35,

Blur: cross-channel; Noise: Salt & Pepper 60% (256 × 256).
Wavelet inpainting

Model: \( \min_x \{ \text{TV}(x) \mid \| P W x - f \| \leq \varepsilon \} \).

Back Projection. 50% data

ADM. SNR: 29.1dB, CPU: 46.2s
MRI reconstruction

Model: \[
\min_x \{ \text{TV}(x) + \tau \| Wx \|_1 \mid \| P\mathcal{F}x - f \| \leq \varepsilon \}.
\]

Left: original (256 × 256); Middle: k-space samples (9.36%, white noise \(10^{-3}\)); Right: recovered by ADM (relative error: 0.48%, CPU: 3s).
Recovery from incomplete data

Model:  \[ \min_x \{ \text{TV}(x) \mid \| PKx - f \| \leq \varepsilon \} \].

Lena: 512 $\times$ 512; Blur: Gaussian, 15$\times$15; White noise: $10^{-3}$; SRs: 5%, 3%, 1%; SNRs: 13.32, 11.98, 6.14dB; CPU: 43, 46, 57s.
Robust linear regression

Given a set of feature vectors $a_i \in \mathbb{R}^n$ and outcomes $b_i \in \mathbb{R}$, $i = 1, \ldots, m$, find weights $x$ that predict the outcome accurately, i.e., $Ax \approx b$. The robust linear regression problem

$$\min_x \|Ax - b\|_1$$

is less sensitive to outliers. Equivalent gauge problem:

$$\min_{x,y}\{\|y\|_1 \mid Ax + y = b\}.$$
Semidefinite programming

The semidefinite programming (SDP) problem

$$\min_X \{ \langle C, X \rangle \mid AX = b, X \succeq 0 \}.$$  

can also be reformulated as gauge optimization. To see this, chose one dual feasible point $y$ so that

$$\hat{C} := C - A^* y \succeq 0.$$  

Then, this SDP is equivalent to a gauge optimization

$$\min_X \{ \langle \hat{C}, X \rangle + \iota_{\geq 0}(X) \mid AX = b \}.$$
Outline

Introduction

Duality
- Lagrange duality
- Fenchel duality
- Gauge duality
Lagrange duality

- Given a general nonlinear programming problem, referred to as primal problem, there exists another closely related nonlinear programming problem, named Lagrange dual problem.
- Under suitable regularity conditions, such as convexity and/or constraint qualifications, the primal and the dual problems have equal optimal objective values.
Lagrange duality

Let $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^q$ and $\mathcal{X} \subseteq \mathbb{R}^n$. The nonlinear constrained optimization problem:

$$ p^* := \inf_{x} \{ f(x) \mid h(x) = 0, g(x) \leq 0, x \in \mathcal{X} \}. $$

The Lagrange function

$$ L(x, \lambda, \mu) = f(x) + \langle \lambda, h(x) \rangle + \langle \mu, g(x) \rangle. $$

The Lagrange dual problem

$$ d^* := \sup_{\lambda, \mu} \{ g(\lambda, \mu) \mid \mu \geq 0 \}, $$

where $g(\lambda, \mu) := \inf_x \{ L(x, \lambda, \mu) \mid x \in X \}$ is the dual objective.
Lagrange duality

- The vectors \( \lambda \) and \( \mu \) are the Lagrange multipliers;
- The Lagrange multipliers corresponding to inequality constraints are restricted to be nonnegative, and those corresponding to equality constraints are unrestricted;
- The dual problem is a concave maximization problem;
- Weak duality \( d^* \leq p^* \);
- Strong duality \( d^* = p^* \) is in general not true. For convex problems (\( X \) convex, \( f \) and \( g \) convex, and \( h \) affine), if Slater’s condition is satisfied, then duality gap vanishes;
- Different Lagrange dual problems can be derived depending on different formulations of the problem.
Example

Consider $\ell_1$-regularized least square problem

$$
\min_x \|x\|_1 + \frac{1}{2}\|Ax - b\|^2.
$$

Two equivalent formulations

$$
\min_{x,z} \left\{ \|x\|_1 + \frac{1}{2}\|Az - b\|^2 \mid x = z \right\},
$$

$$
\min_{x,z} \left\{ \|x\|_1 + \frac{1}{2}\|z - b\|^2 \mid Ax = z \right\}.
$$

Their Lagrange dual problems

$$
\max_{\|y\|_\infty \leq 1} \min_z \langle y, z \rangle + \frac{1}{2}\|Az - b\|^2,
$$

$$
\max_{\|A^T y\|_\infty \leq 1} -\langle b, y \rangle - \frac{1}{2}\|y\|^2.
$$
Fenchel duality

Let $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$. Consider unconstrained optimization

$$\min_x f(x).$$

Any constraints must be incorporated into the objective function via the indicator function.

The key ingredient of Fenchel duality is that any closed convex function is the pointwise supremum of the set of affine functions which minorize it. This leads to the Fenchel conjugate function:

$$f^*(y) = \sup_x \{ \langle y, x \rangle - f(x) \}, \forall y.$$
Fenchel duality

Assume that $\phi(x, y)$ is a perturbation function of $f(x)$, i.e.,

$$\phi(x, 0) = f(x), \quad \forall x,$$

and define

$$h(y) := \inf_x \phi(x, y), \quad \forall y.$$  

Then, minimizing $f$ is equivalent to evaluating $h(0)$. Can show that

$$- \inf_y \phi^*(0, y) = h^{**}(0) \leq h(0) = \inf_x \phi(x, 0).$$

The problem of evaluating $h^{**}(0)$ is called the Fenchel dual problem. Strong duality $h^{**}(0) = h(0)$ holds if and only if

$$\inf_x \sup_y \{\langle x, y \rangle - h(y)\} = \sup_y \inf_x \{\langle x, y \rangle - h(y)\}.$$
Fenchel duality

Theorem (Fenchel duality theorem)

Let \( f : \mathbb{R}^n \to \mathbb{R} \cup \{ +\infty \} \) and \( -g : \mathbb{R}^n \to \mathbb{R} \cup \{ +\infty \} \) be closed proper convex functions. Then, under regularity conditions it holds that

\[
\inf_x \{ f(x) - g(x) \} = \sup_y \{ g^*(y) - f^*(y) \},
\]

where \( f^* \) is the convex conjugate of \( f \) and \( g^* \) is the concave conjugate of \( g \), i.e.,

\[
f^*(y) := \sup_x \{ \langle y, x \rangle - f(x) \},
\]

\[
g^*(y) := \inf_x \{ \langle y, x \rangle - g(x) \}.
\]
Geometric interpretation

\[ [\text{slop, dual obj}] = [-2.05, -109.28] \]
Geometric interpretation

\[[\text{slop}, \text{dual obj}] = [-1.60, -73.00]\]
Geometric interpretation

[slop, dual obj] = [-1.10, -38.62]
Geometric interpretation

\[ \text{slop, dual obj} = [-0.60, -10.50] \]
Geometric interpretation

\[ [\text{slop, dual obj}] = [-0.10, 11.38] \]
Geometric interpretation

\[ \text{slop, dual obj} = [1.40, 39.50] \]
Geometric interpretation

\[ \text{slop, dual obj} = [1.90, 36.38] \]
Geometric interpretation

\[ [\text{slop, dual obj}] = [2.40, 27.00] \]
Gauge optimization

Definition (gauge optimization)
Minimizing a closed gauge function $\kappa$ over a closed convex set $C$ is called a gauge optimization. That is

$$\min_{x} \{ \kappa(x) \mid x \in C \}. \quad \text{(GP)}$$

Definition (feasible point)

- A point $x \in C$ is called a feasible point of GP;
- If $x \in C$ and $\kappa(x) = +\infty$, then $x$ is essentially infeasible;
- If $x \in C$ and $\kappa(x) < +\infty$, then $x$ is strongly feasible;
- Similar definitions for the gauge dual problem defined below.
Gauge optimization

**Definition (antipolar set)**
The antipolar of a closed convex set $C$ is given by

$$C' = \{ y \in \mathbb{R}^n \mid \langle x, y \rangle \geq 1 \text{ for all } x \in C \}.$$ 

**Definition (polar function)**
The polar function of a gauge $\kappa$ is defined as

$$\kappa^\circ(y) = \sup \{ \langle x, y \rangle \mid \kappa(x) \leq 1 \}.$$ 

In particular, the polar function of a norm is its dual norm.

**Definition (gauge dual)**
The nonlinear gauge dual problem of (GP) is defined by

$$\min_y \{ \kappa^\circ(y) \mid y \in C' \}.$$  \hfill (GD)
Theorem (weak duality, Freund 1987)

Let $p^*$ and $d^*$ be respectively the optimal values for GP and GD.

1. If $x$ and $y$ are, resp., strongly feasible for GP and GD, then $\kappa(x)\kappa^\circ(y) \geq 1$, and hence $p^*d^* \geq 1$.
2. If $p^* = 0$, then $d^* = +\infty$, i.e., GD is essentially infeasible;
3. If $d^* = 0$, then $p^* = +\infty$, i.e., GP is essentially infeasible.
Theorem (strong duality, Freund 1987)

Let $p^*$ and $d^*$ be respectively the optimal values for GP and GD. If

$$\text{(ri } C\text{)} \cap (\text{ri dom } \kappa) \neq \emptyset$$

and

$$\text{(ri } C'\text{)} \cap (\text{ri dom } \kappa^\circ) \neq \emptyset,$$

then $p^* d^* = 1$, and each program attains its optimum.
Gauge duality

The Fenchel dual of GP is given by

$$d_f^* = \max_y \{-\nu^*_C(-y) \mid k^\circ(y) \leq 1\}. \quad \text{(FD)}$$

Theorem (weak duality, Friedlander et al., 2014)

Suppose that $\text{dom } k^\circ \cap C' \neq \emptyset$. Then

$$\mu^* \geq d_f^* = 1/d^* > 0.$$ 

Furthermore,

1. if $y^*$ solves FD, then $y^*/d_f^*$ solves GD;
2. if $y^*$ solves GD and $d^* > 0$, then $y^*/d^*$ solves FD.
Theorem (strong duality, Friedlander et al., 2014)

1. Suppose that

\[ \text{dom } k^\circ \cap C' \neq \emptyset \text{ and } (\text{ri dom } \kappa) \cap (\text{ri } C) \neq \emptyset, \]

then \( p^*d^* = 1 \) and GD attains its optimal value.

2. Suppose that

\[ (\text{ri dom } \kappa) \cap (\text{ri } C) \neq \emptyset \text{ and } (\text{ri dom } \kappa^\circ) \cap (\text{ri } C') \neq \emptyset, \]

then \( p^*d^* = 1 \) and both GP and GD attain their optimal values.
Gauge duality

Consider the gauge optimization

$$\min_x \{ \kappa(x) \mid \rho(Ax - b) \leq \varepsilon \}.$$ 

Lagrange dual

$$\max_y \{ b^T y - \varepsilon \rho^\circ(y) \mid \kappa^\circ(A^T y) \leq 1 \}$$ 

Gauge dual

$$\min_y \{ \kappa^\circ(A^T y) \mid b^T y - \varepsilon \rho^\circ(y) \geq 1 \}$$ 

In general, problems with simper constraints and complex objective are more preferred than those with simpler objective yet complex constraints.
Acknowledgements

Some pictures are taken from Michael P. Friedlander’s SIAM Optimization talk in 2011.