Dual and primal-dual methods

Yuxin Chen
Princeton University, Spring 2018
Outline

- Dual proximal gradient method
- Primal-dual proximal gradient method
Dual proximal gradient method
Constrained convex optimization

\[
\begin{align*}
\text{minimize}_x & \quad f(x) \\
\text{subject to} & \quad Ax + b \in C
\end{align*}
\]

where \( f \) is convex, and \( C \) is convex set

- projection onto feasible set could be highly nontrivial (even when projection onto \( C \) is easy)
More generally,

$$\minimize_x f(x) + h(Ax)$$

where $f$ and $h$ are convex

- proximal operator w.r.t. $\tilde{h}(x) := h(Ax)$ could be difficult (even when $\text{prox}_h$ is inexpensive)
A possible route: dual formulation

\[ \text{minimize}_x \quad f(x) + h(Ax) \]

\[ \uparrow \quad \text{add auxiliary variable } z \]

\[ \text{minimize}_{x,z} \quad f(x) + h(z) \]

subject to \[ Ax = z \]

dual formulation:

\[ \text{maximize}_\lambda \quad \min_{x,z} \left[ f(x) + h(z) + \langle \lambda, A^T(x - z) \rangle \right] \]

\[ := L(x,z,\lambda) \quad (\text{Lagrangian}) \]
A possible route: dual formulation

maximize
\[ \lambda \]
min
\[ x, z \]
\[ f(x) + h(z) + \langle \lambda, Ax - z \rangle \]

\[ \uparrow \text{decouple } x \text{ and } z \]

maximize
\[ \lambda \]
min
\[ x, z \]
\[ \{ \langle A^\top \lambda, x \rangle + f(x) \} + \min_z \{ h(z) - \langle \lambda, z \rangle \} \]

\[ \uparrow \]

maximize
\[ \lambda \]
\[ f^*(-A^\top \lambda) - h^*(\lambda) \]

where \( f^* \) (resp. \( h^* \)) is Fenchel conjugate of \( f \) (resp. \( h \))
Primal vs. dual problems

\[(\text{primal}) \quad \text{minimize}_x \quad f(x) + h(Ax)\]

\[(\text{dual}) \quad \text{minimize}_\lambda \quad f^*(-A^\top \lambda) + h^*(\lambda)\]

Dual formulation is useful if

- proximal operator w.r.t. $h$ is cheap (recall Moreau decomposition $\text{prox}_h(x) + \text{prox}_{h^*}(x) = x$)
- $f^*$ is smooth (or if $f$ is strongly convex)
Dual proximal gradient methods

Apply proximal gradient methods for dual problem:

**Algorithm 9.1 Dual proximal gradient algorithm**

1. **for** $t = 0, 1, \cdots$ **do**
2. $\lambda^{t+1} = \text{prox}_{\eta_t h^*}(\lambda^t + \eta_t A \nabla f^*(-A^\top \lambda^t))$

- let $Q(\lambda) := -f^*(-A^\top \lambda) - h^*(\lambda)$ and $Q^{\text{opt}} = \max_{\lambda} Q(\lambda)$, then
  $Q^{\text{opt}} - Q(\lambda^t) \lesssim \frac{1}{t}$
Algorithm 9.2 Dual proximal gradient algorithm (primal representation)

1: for $t = 0, 1, \cdots$ do
2: $x^t = \arg \min_x \{ f(x) + \langle A^\top \lambda^t, x \rangle \}$
3: $\lambda^{t+1} = \lambda^t + \eta_t A x^t - \eta_t \text{prox}_{\eta_t^{-1} h}(\eta_t^{-1} \lambda^t + A x^t)$

- $\{x^t\}$ is primal sequence, which is nevertheless not always feasible
By definition of $x^t$,

$$-A^\top \lambda^t \in \partial f(x^t)$$

This together with conjugate subgradient theorem and smoothness of $f^*$ yields

$$x^t = \nabla f^*(-A^\top \lambda^t)$$

Therefore, dual proximal gradient update rule can be rewritten as

$$\lambda^{t+1} = \text{prox}_{\eta_t h^*}(\lambda^t + \eta_t A x^t)$$
Moreover, from extended Moreau decomposition, we know

$$\text{prox}_{\eta_t h^*}(\lambda^t + \eta_t A x^t) = \lambda^t + \eta_t A x^t - \eta_t \text{prox}_{\eta_t^{-1} h}(\eta_t^{-1} \lambda^t + A x^t)$$

$$\implies \lambda^{t+1} = \lambda^t + \eta_t A x^t - \eta_t \text{prox}_{\eta_t^{-1} h}(\eta_t^{-1} \lambda^t + A x^t)$$
Accuracy of primal sequence

One can control primal accuracy via dual accuracy:

**Lemma 9.1**

Let $x_\lambda := \arg \min_x \{ f(x) + \langle A^\top \lambda, x \rangle \}$. Suppose $f$ is $\mu$-strongly convex. Then

$$\|x^* - x_\lambda\|_2^2 \leq \frac{2(Q^{\text{opt}} - Q(\lambda))}{\mu}$$

- **consequence:** $\|x^* - x^t\|_2^2 \lesssim 1/t$
Proof of Lemma 9.1

Recall that Lagrangian is given by

\[ \mathcal{L}(x, z, \lambda) := f(x) + \langle A^\top \lambda, x \rangle + h(z) - \langle \lambda, z \rangle \]

\[ := \tilde{f}(x, \lambda) \quad \text{and} \quad := \tilde{h}(z, \lambda) \]

For any \( \lambda \), define \( x_\lambda := \arg \min_x \tilde{f}(x, \lambda) \) and \( z_\lambda := \arg \min_z \tilde{h}(x, \lambda) \) (non-rigorous). Then by strong convexity,

\[ \mathcal{L}(x^*, z^*, \lambda) - \mathcal{L}(x_\lambda, z_\lambda, \lambda) \geq \tilde{f}(x^*, \lambda) - \tilde{f}(x_\lambda, \lambda) \geq \frac{1}{2} \mu \|x^* - x_\lambda\|^2 \]

In addition, since \( Ax^* = z^* \), one has

\[ \mathcal{L}(x^*, z^*, \lambda) = f(x^*) + h(z^*) + \langle \lambda, Ax^* - z^* \rangle = f(x^*) + h(Ax^*) = F_{\text{opt}} \]

\[ = Q_{\text{opt}} \]

This combined with \( \mathcal{L}(x_\lambda, z_\lambda, \lambda) = Q(\lambda) \) gives

\[ Q_{\text{opt}} - Q(\lambda) \geq \frac{1}{2} \mu \|x^* - x_\lambda\|^2 \]

as claimed.

Dual and primal-dual method
Accelerated dual proximal gradient methods

One can apply FISTA to dual problem to improve convergence:

**Algorithm 9.3** Accelerated dual proximal gradient algorithm

1: for $t = 0, 1, \ldots$ do
2: $\lambda^{t+1} = \text{prox}_{\eta_t h^*}(w^t + \eta_t A \nabla f^*( - A^T w^t))$
3: $\theta_{t+1} = \frac{1 + \sqrt{1 + 4\theta_t^2}}{2}$
4: $w^{t+1} = \lambda^{t+1} + \frac{\theta_{t-1}}{\theta_{t+1}}(\lambda^{t+1} - \lambda^t)$

- apply FISTA theory and Lemma 9.1 to get

$$Q^{\text{opt}} - Q(\lambda^t) \lesssim \frac{1}{t^2} \quad \text{and} \quad ||x^* - x^t||_2^2 \lesssim \frac{1}{t^2}$$
Primal representation of accelerated dual proximal gradient methods

Algorithm 9.3 admits more explicit primal representation

Algorithm 9.4 Accelerated dual proximal gradient algorithm (primal representation)

1: \textbf{for} \ t = 0, 1, \cdots \ \textbf{do}
2: \ \ \ x^t = \text{arg} \min_x f(x) + \langle A^\top w^t, x \rangle
3: \ \ \ \lambda^{t+1} = w^t + \eta_t A x^t - \eta_t \text{prox}_{\eta_t^{-1} h}(\eta_t^{-1} w^t + A x^t)
4: \ \ \ \theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}
5: \ \ \ w^{t+1} = \lambda^{t+1} + \frac{\theta_t-1}{\theta_{t+1}} (\lambda^{t+1} - \lambda^t)
Primal-dual proximal gradient method
Nonsmooth optimization

\[
\text{minimize}_x \quad f(x) + h(Ax)
\]

where \( f \) and \( h \) are closed and convex

- both \( f \) and \( h \) might be non-smooth
- both \( f \) and \( h \) admit inexpensive proximal operators
Primal-dual approaches?

\[
\text{minimize}_x \quad f(x) + h(Ax)
\]

So far we have discussed proximal method (resp. dual proximal method), which essentially updates only primal (resp. dual) variables.

**Question:** can we update both primal and dual variables simultaneously and take advantage of both \(\text{prox}_f\) and \(\text{prox}_h\)?
To this end, we first derive saddle-point formulation that includes both primal and dual variables

\[
\begin{align*}
\text{minimize}_x & \quad f(x) + h(Ax) \\
\iff & \quad \text{add auxiliary variable } z \\
\text{minimize}_{x,z} & \quad f(x) + h(z) \quad \text{subject to } Ax = z \\
\iff & \\
\text{maximize}_\lambda \min_{x,z} & \quad f(x) + h(z) + \langle \lambda, Ax - z \rangle \\
\iff & \\
\text{maximize}_\lambda \min_x & \quad f(x) + \langle \lambda, Ax \rangle - h^*(\lambda) \\
\iff & \\
\text{minimize}_x \max_\lambda & \quad f(x) + \langle \lambda, Ax \rangle - h^*(\lambda) \quad \text{(saddle-point problem)}
\end{align*}
\]
A saddle-point formulation

\[ \text{minimize}_{\mathbf{x}} \max_{\lambda} f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} \rangle - h^*(\lambda) \]  \quad (9.1)

- one can then consider updating primal variable \( \mathbf{x} \) and dual variable \( \lambda \) simultaneously
- we’ll first examine optimality condition for (9.1), which in turn gives ideas about how to jointly update primal and dual variables
Optimality condition

\[
\text{minimize}_x \ \max_\lambda \ f(x) + \langle \lambda, Ax \rangle - h^*(\lambda)
\]

optimality condition:

\[
\begin{cases}
0 \in \partial f(x) + A^\top \lambda \\
0 \in Ax - \partial h^*(\lambda)
\end{cases}
\]

\[
\iff \quad 0 \in \begin{bmatrix}
-A & A^\top \\
\end{bmatrix} \begin{bmatrix}
x \\
\lambda
\end{bmatrix} + \begin{bmatrix}
\partial f(x) \\
\partial h^*(\lambda)
\end{bmatrix} := \mathcal{F}(x, \lambda) \quad (9.2)
\]

key idea: iteratively update \((x, \lambda)\) to reach a point obeying \(0 \in \mathcal{F}(x, \lambda)\)
How to solve $0 \in \mathcal{F}(x)$ in general?

In general, finding solution to

$$0 \in \mathcal{F}(x)$$

called "monotone inclusion problem" if $\mathcal{F}$ is maximal monotone

$$\iff x \in (\mathcal{I} + \mathcal{F})(x)$$

is equivalent to finding fixed point of $(\mathcal{I} + \eta \mathcal{F})^{-1}$, i.e. solution to resolvent of $\mathcal{F}$

$$x = (\mathcal{I} + \eta \mathcal{F})^{-1}(x)$$

This suggests natural fixed-point iteration / resolvent iteration:

$$x^{t+1} = (\mathcal{I} + \eta \mathcal{F})^{-1}(x^t), \quad t = 0, 1, \cdots$$
Aside: monotone operators

— Ryu, Boyd ’16

For a CCP functions, strong convexity and strong smoothness are dual properties; a CCP $f$ is strongly convex with parameter $m$ if and only if $f^*$ is strongly smooth with parameter $L = 1/m$, and vice versa. We discuss these claims in the appendix.

For example, $f(x) = x^2/2 + |x|$, where $x \in \mathbb{R}$, is strongly convex with parameter 1 but not strongly smooth. Its conjugate is $f^*(x) = (|x| - 1)^+ + x^2/2$, where $(\cdot)^+$ denotes the positive part, and is strongly smooth with parameter 1 but not strongly convex. See Fig. 4.

Figure 4. Example of $f$ and its conjugate $f^*$.

4.3. Examples

Relations on $\mathbb{R}$. We describe this informally. A relation on $\mathbb{R}$ is monotone if it is a curve in $\mathbb{R}^2$ that is always nondecreasing; it can have horizontal (flat) portions and also vertical (infinite slope) portions. If it is a continuous curve with no end points, then it is maximal monotone. It is strongly monotone with parameter $m$ if it maintains a minimum slope $m$ everywhere; it has Lipschitz constant $L$ if its slope is never more than $L$. See Fig. 5.

Figure 5. Examples of operators on $\mathbb{R}$.

Continuous functions. A continuous monotone function $F : \mathbb{R}^n \to \mathbb{R}^n$ (with $\text{dom} F = \mathbb{R}^n$) is maximal. Let us show this. Assume for contradiction that there is a pair $(\tilde{x}, \tilde{u}) \notin F$, such that $(\tilde{u} - F(x))^T(\tilde{x} - x) \geq 0$.

• relation $F$ is called monotone if

$$\langle u - v, x - y \rangle \geq 0, \quad \forall (x, u), (y, v) \in F$$

• relation $F$ is called maximal monotone if there is no monotone operator that contains it
Proximal point method

\[ \mathbf{x}^{t+1} = (\mathcal{I} + \eta_t \mathcal{F})^{-1}(\mathbf{x}^t), \quad t = 0, 1, \ldots \]

If \( \mathcal{F} = \partial f \) for some convex function \( f \), then this proximal point method becomes

\[ \mathbf{x}^{t+1} = \text{prox}_{\eta_t f}(\mathbf{x}^t), \quad t = 0, 1, \ldots \]

- useful when \( \text{prox}_{\eta_t f} \) is cheap
Recall that we want to solve

\[ 0 \in \begin{bmatrix} -A & A^\top \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} + \begin{bmatrix} \partial f(x) \\ \partial h^*(\lambda) \end{bmatrix} := \mathcal{F}(x, \lambda) \]

**issue of proximal point method:** computing \((\mathcal{I} + \eta \mathcal{F})^{-1}\) is in general difficult
Back to primal-dual approaches

**observation:** practically we may often consider splitting $\mathcal{F}$ into two operators

$$0 \in \mathcal{A}(x, \lambda) + \mathcal{B}(x, \lambda)$$

with

$$\mathcal{A}(x, \lambda) = \begin{bmatrix} A & x \\ -A^\top & \lambda \end{bmatrix}, \quad \mathcal{B}(x, \lambda) = \begin{bmatrix} \partial f(x) \\ \partial h^*(\lambda) \end{bmatrix}$$

(9.3)

- $(I + \eta A)^{-1}$ can be computed by solving linear systems
- $(I + \eta B)^{-1}$ is easy if prox$_f$ and prox$_{h^*}$ are both inexpensive

**solution:** design update rules based on $(I + \eta A)^{-1}$ and $(I + \eta B)^{-1}$ instead of $(I + \eta F)^{-1}$
We now introduce one principled approach based on operator splitting

\[ \text{find } x \text{ s.t. } 0 \in \mathcal{F}(x) = A(x) + B(x) \]

operator splitting

let \( \mathcal{R}_A := (\mathcal{I} + \eta A)^{-1} \) and \( \mathcal{R}_B := (\mathcal{I} + \eta B)^{-1} \) be resolvents, and
\( C_A := 2\mathcal{R}_A - \mathcal{I} \) and \( C_B := 2\mathcal{R}_B - \mathcal{I} \) be Cayley operators

**Lemma 9.2**

\[ 0 \in A(x) + B(x) \iff C_A C_B(z) = z \text{ with } x = \mathcal{R}_B(z) \quad (9.4) \]

it comes down to finding fixed point of \( C_A C_B \)
Operator splitting via Cayley operators

\[ x \in \mathcal{R}_{A+B}(x) \iff C_A C_B(z) = z \]

- **advantage:** allows to apply \( C_A \) (resp. \( \mathcal{R}_A \)) and \( C_B \) (resp. \( \mathcal{R}_B \)) sequentially (instead of computing \( \mathcal{R}_{A+B} \))
Proof of Lemma 9.2

\[ C_A C_B(z) = z \]

\[ x = \mathcal{R}_B(z) \quad (9.5a) \]

\[ \iff \tilde{z} = 2x - z \quad (9.5b) \]

\[ \tilde{x} = \mathcal{R}_B(\tilde{z}) \quad (9.5c) \]

\[ z = 2\tilde{x} - \tilde{z} \quad (9.5d) \]

From (9.5b) and (9.5d), we see that

\[ \tilde{x} = x \]

which together with (9.5d) gives

\[ 2x = z + \tilde{z} \quad (9.6) \]
Proof of Lemma 9.2 (cont.)

Recall that

\[ z \in x + \eta B(x) \quad \text{and} \quad \tilde{z} \in x + \eta A(x) \]

Adding these two facts and using (9.6), we get

\[ 2x = z + \tilde{z} \in 2x + \eta B(x) + \eta A(x) \]

\[ \iff \quad 0 \in A(x) + B(x) \]
How to find points obeying $x = C_A C_B(x)$?

- First attempt: fixed-point iteration
  \[ z^{t+1} = C_A C_B(z^t) \]
  unfortunately, it may not converge in general

- **Douglas-Rachford splitting**: damped fixed-point iteration
  \[ z^{t+1} = \frac{1}{2} (I + C_A C_B)(z^t) \]
  converges when solution to $0 \in A(x) + B(x)$ exists!
Douglas-Rachford splitting update rule $z^{t+1} = \frac{1}{2}(\mathcal{I} + C_A C_B)(z^t)$ is essentially:

\begin{align*}
  x^{t+\frac{1}{2}} &= \mathcal{R}_B(z^t) \\
  z^{t+\frac{1}{2}} &= 2x^{t+\frac{1}{2}} - z^t \\
  x^{t+1} &= \mathcal{R}_A(z^{t+\frac{1}{2}}) \\
  z^{t+1} &= \frac{1}{2}(z^t + 2x^{t+1} - z^{t+\frac{1}{2}}) \\
  &= z^t + x^{t+1} - x^{t+\frac{1}{2}}
\end{align*}

where $x^{t+\frac{1}{2}}$ and $z^{t+\frac{1}{2}}$ are auxiliary variables.
More explicit expression for D-R splitting

or equivalently,

\[ x^{t+\frac{1}{2}} = \mathcal{R}_B(z^t) \]
\[ x^{t+1} = \mathcal{R}_A(2x^{t+\frac{1}{2}} - z^t) \]
\[ z^{t+1} = z^t + x^{t+1} - x^{t+\frac{1}{2}} \]
Douglas-Rachford primal-dual splitting

\[
\text{minimize}_x \ \max_\lambda \ f(x) + \langle \lambda, Ax \rangle - h^*(\lambda)
\]

Applying Douglas-Rachford splitting to (9.3) yields

\[
x_{t+\frac{1}{2}} = \text{prox}_{\eta f}(p^t)
\]

\[
\lambda_{t+\frac{1}{2}} = \text{prox}_{\eta g^*}(q^t)
\]

\[
\begin{bmatrix}
x^{t+1} \\
\lambda^{t+1}
\end{bmatrix} = \begin{bmatrix}
I & \eta A^T \\
-\eta A & I
\end{bmatrix}^{-1} \begin{bmatrix}
2x_{t+\frac{1}{2}} - p^t \\
2\lambda_{t+\frac{1}{2}} - q^t
\end{bmatrix}
\]

\[
p^{t+1} = p^t + x^{t+1} - x^{t+\frac{1}{2}}
\]

\[
q^{t+1} = q^t + \lambda^{t+1} - \lambda^{t+\frac{1}{2}}
\]
Example

\[
\begin{align*}
\text{minimize}_x & \quad \|x\|_2 + \gamma \|Ax - b\|_1 \\
\iff \quad & \text{minimize}_x \quad f(x) + g(Ax) \\
\text{with } f(x) & := \|x\|_2 \quad \text{and } g(y) := \gamma \|y - b\|_1
\end{align*}
\]

— Connor, Vandenberghe ’14
Example

\[
\begin{align*}
\text{minimize} & \quad \|Kx - b\|_1 + \gamma \|Dx\|_{\text{iso}} \\
\text{s.t.} & \quad 0 \leq x \leq 1
\end{align*}
\]

\[\Leftrightarrow \quad \text{minimize}_x \quad f(x) + g(Ax)\]

with \(f(x) := 1_{\{0 \leq x \leq 1\}}(x)\) and \(g(y_1, y_2) := \|y_1 - b\|_1 + \gamma \|y_2\|_{\text{iso}}\)

![Graph showing convergence plots]

---

— Connor, Vandenberghe ’14
Reference

[1] "Optimization methods for large-scale systems, EE236C lecture notes," L. Vandenberghe, UCLA.


