Projection onto $Ax \leq b$ by dual descent

Andersen Ang
U.Southampton UK
angms.science
May 31, 2025

1st ver. May 23, 2025

Content

Contents

Problem Setup $(\mathcal{P}) : \underset{\boldsymbol{x}}{\operatorname{argmin}} \ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 \ \text{ s.t. } \ \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}.$

Lagrangian $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\rangle$ and dual $d(\boldsymbol{\lambda}) = -\frac{1}{2}\|\boldsymbol{A}^\top\boldsymbol{\lambda}\|_2^2 + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{u}-\boldsymbol{b}\rangle$

Solving the dual problem $(\mathcal{D}) : \underset{\boldsymbol{\lambda}\geq \mathbf{0}}{\operatorname{argmin}} \ d(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{A}^\top\boldsymbol{\lambda}\|_2^2 - \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{u}-\boldsymbol{b}\rangle$

Hoffman's Lemma and geometric sensitivity

Problem setup

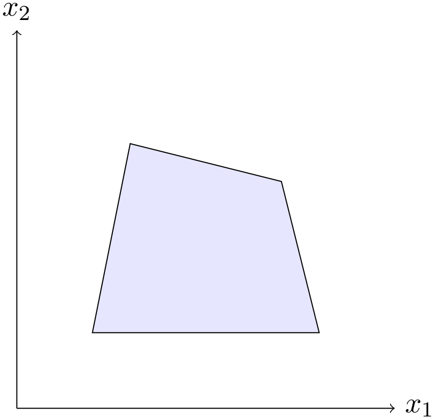
$$(\mathcal{P}) : \operatorname*{argmin}_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 \text{ s.t. } \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}.$$

- $\boldsymbol{x}, \boldsymbol{u}$ are points, $\boldsymbol{u} \in \mathbb{R}^n$ is given, $\boldsymbol{x} \in \mathbb{R}^n$ is the variable
- $\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2$ is the distance between $\boldsymbol{x}$ and $\boldsymbol{u}$
- $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ is a constraint
  - $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is a given matrix
  - $\boldsymbol{b} \in \mathbb{R}^m$ is a given vector
  - $\leq$ is applied element-wise
  - we make no assumption on $m \geq n$ nor $n \geq m$

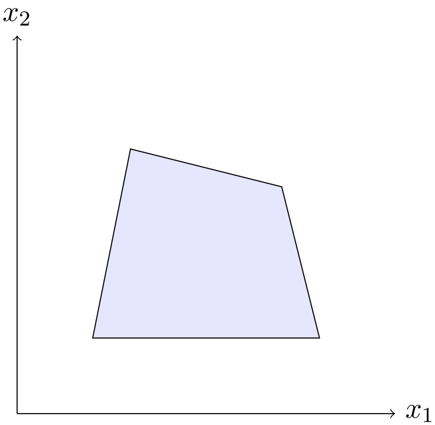
Solving $\underset{\boldsymbol{x}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2$ s.t. $\boldsymbol{Ax} \le \boldsymbol{b}$ means projecting u onto a set

- We are giving a point $\boldsymbol{u} \in \mathbb{R}^n$
- We want to project $\boldsymbol{u}$ onto C
- Solving $\underset{\boldsymbol{x}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2$ s.t. $\boldsymbol{Ax} \le \boldsymbol{b}$ means projecting $\boldsymbol{u}$ onto C
- If $\boldsymbol{u} \in C$, the projection does nothing and return $\boldsymbol{u}$
- If $\boldsymbol{u} \notin C$, the projection return a point in C that is closest to $\boldsymbol{u}$
- Projection onto a polygon / polytope is generally hard

An example of C described by $\boldsymbol{Ax} \le \boldsymbol{b}$



Projecting onto a set is not easy in general



https://angms.science/doc/CVX/POCS.pdf

Contents

Lagrangian

$$(\mathcal{P}) : \operatorname*{\mathbf{argmin}}_{\boldsymbol{x}} \ \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \ \text{ s.t. } \ \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$$

- The Lagrangian function is
$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \rangle$$
$\boldsymbol{\lambda} \in \mathbb{R}^m$ is the Lagrangian multiplier

- We have
$$\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \boldsymbol{x} - \boldsymbol{u} + \boldsymbol{A}^\top \boldsymbol{\lambda}$$

- The KKT system of $\mathcal{P}$ is
$$\begin{cases} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \mathbf{0} & \text{Stationarity} \\ \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} & \text{Primal feasibility} \\ \boldsymbol{\lambda} \geq \mathbf{0} & \text{Dual feasibility} \\ \boldsymbol{\lambda} \perp \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} & \text{Complementary slackness} \end{cases}$$

- Suppose $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ solves the KKT system, then $\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \mathbf{0}$ gives $\boldsymbol{x}^* = \boldsymbol{u} - \boldsymbol{A}^\top \boldsymbol{\lambda}^*$

Driving the dual

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \rangle$$

- Dual $d(\boldsymbol{\lambda}) = \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$

- By $\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \mathbf{0}$ gives $\boldsymbol{x} = \boldsymbol{u} - \boldsymbol{A}^\top \boldsymbol{\lambda}$, we have

$$\begin{aligned} d(\boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{u} - \boldsymbol{A}^\top \boldsymbol{\lambda}, \boldsymbol{\lambda}) &= \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{A}^\top \boldsymbol{\lambda} - \boldsymbol{u}\|_2^2 + \langle \boldsymbol{\lambda}, \boldsymbol{A}(\boldsymbol{u} - \boldsymbol{A}^\top \boldsymbol{\lambda}) - \boldsymbol{b} \rangle \\ &= \frac{1}{2}\|\boldsymbol{A}^\top \boldsymbol{\lambda}\|_2^2 + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{u} - \boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{\lambda} - \boldsymbol{b} \rangle \\ &= -\frac{1}{2}\|\boldsymbol{A}^\top \boldsymbol{\lambda}\|_2^2 + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b} \rangle \end{aligned}$$

- Dual problem

$$\begin{aligned} \operatorname*{argmax}_{\boldsymbol{\lambda} \geq \mathbf{0}} d(\boldsymbol{\lambda}) &= \operatorname*{argmax}_{\boldsymbol{\lambda} \geq \mathbf{0}} -\frac{1}{2}\|\boldsymbol{A}^\top \boldsymbol{\lambda}\|_2^2 + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b} \rangle \\ &= \operatorname*{argmin}_{\boldsymbol{\lambda} \geq \mathbf{0}} \frac{1}{2}\|\boldsymbol{A}^\top \boldsymbol{\lambda}\|_2^2 - \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b} \rangle \end{aligned}$$

A dual-based method, summary

Goal: solve $$(\mathcal{P}) : \underset{\boldsymbol{x}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 \text{ s.t. } \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$$

Lagrangian $$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\rangle$$

Dual problem $$(\mathcal{D}) : \underset{\boldsymbol{\lambda} \geq \boldsymbol{0}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{A}^\top \boldsymbol{\lambda}\|_2^2 - \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}\rangle$$

Get $\boldsymbol{\lambda}^*$ form $\mathcal{D}$, then $$\boldsymbol{x}^* = \boldsymbol{u} - \boldsymbol{A}^\top \boldsymbol{\lambda}^*$$

Contents

Solving the dual problem by projected gradient descent

$$(\mathcal{D}) : \operatorname*{argmin}_{\boldsymbol{\lambda} \geq \mathbf{0}} d(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{A}^\top \boldsymbol{\lambda}\|_2^2 - \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b} \rangle$$

- Projected gradient descent $\boldsymbol{\lambda}_{k+1} = \Big[\boldsymbol{\lambda}_k - \alpha \nabla d(\boldsymbol{\lambda}_k)\Big]_+$

- Gradient $\nabla d(\boldsymbol{\lambda}) = \boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{\lambda} - (\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b})$

- Stepsize, by the theory of gradient descent $\alpha = \dfrac{1}{\|\boldsymbol{A}\boldsymbol{A}^\top\|_2}$

- Compute $\boldsymbol{Q} = \boldsymbol{A}\boldsymbol{A}^\top$, $\boldsymbol{p} = \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}$, $\|\boldsymbol{Q}\|_2$, then

$$\boldsymbol{\lambda}_{k+1} \;=\; \left[\boldsymbol{\lambda}_k - \frac{\boldsymbol{Q}\boldsymbol{\lambda} - \boldsymbol{p}}{\|\boldsymbol{Q}\|_2}\right]_+ \;=\; \left[\left(\boldsymbol{I}_m - \frac{\boldsymbol{Q}}{\|\boldsymbol{Q}\|_2}\right)\boldsymbol{\lambda}_k + \frac{\boldsymbol{Q}\boldsymbol{p}}{\|\boldsymbol{Q}\|_2}\right]_+ \tag{ProjGD-Dual}$$

Algorithm 1: ProjGD-Dual

Input: $\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\lambda}_0, \boldsymbol{u}$

$\boldsymbol{Q} = \boldsymbol{A}\boldsymbol{A}^\top$, $\boldsymbol{p} = \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}$, $\|\boldsymbol{Q}\|_2$ `// pre-compute constant`

for $k = 1, 2, \ldots$ **do**

$\quad \boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \dfrac{\boldsymbol{Q}\boldsymbol{\lambda}_k - \boldsymbol{p}}{\|\boldsymbol{Q}\|_2}\right]_+$ `// projected gradient descent step`

Algorithm 2: ProjGD-Dual + Nesterov acceleration

Input: $\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\lambda}_0, \boldsymbol{u}$

$\boldsymbol{Q} = \boldsymbol{A}\boldsymbol{A}^\top$, $\boldsymbol{p} = \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}$, $\|\boldsymbol{Q}\|_2$, $\boldsymbol{y}_0 = \boldsymbol{\lambda}_0$, $t_0 = 1$ `// pre-compute constant`

for $k = 1, 2, \ldots$ **do**

$\quad \boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \dfrac{\boldsymbol{Q}\boldsymbol{y}_k - \boldsymbol{p}}{\|\boldsymbol{Q}\|_2}\right]_+$ `// projected gradient descent step`

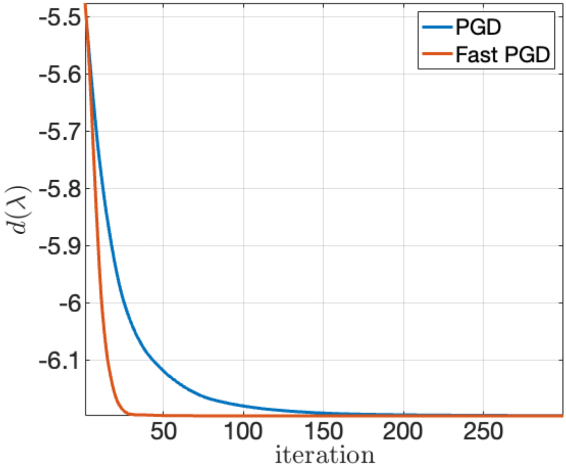
$\quad t_{k+1} = \dfrac{1 + \sqrt{1 + 4t_k^2}}{2}$ `// momentum parameter`

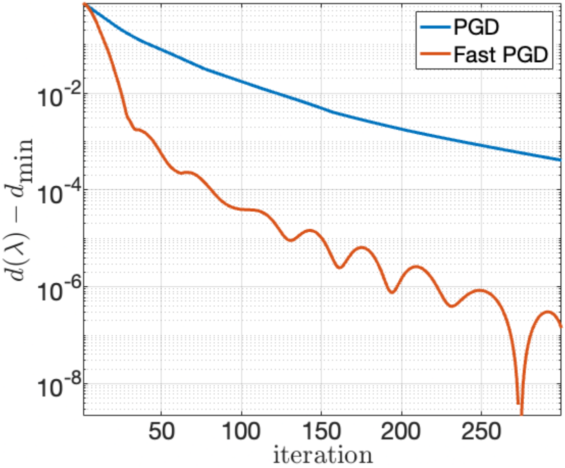
$\quad \boldsymbol{y}_{k+1} = \boldsymbol{\lambda}_{k+1} + \dfrac{t_k - 1}{t_{k+1}}(\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k)$ `// Nesterov acceleration`

```
clear;close all;clc
m = 30;   n = 50;
A = rand(m,n);  b   = rand(m,1);    u = rand(n,1);
Q = A*A';       nQ2 = norm(Q,2);    p = A*u-b;
max_iter = 1000;
lambda_ini = zeros(m, 1);   % Initial dual variable (must be >= 0)

%% Projected gradient descent
lambda = lambda_ini;
for k = 1:max_iter
  lambda = max(lambda - (Q * lambda - p)/nQ2, 0);
d(k) = 0.5*norm(A'*lambda)^2 - lambda'*p;
end

%% + Nesterov
lambda = lambda_ini;
y = lambda;                    % Momentum term
t = 1;                         % Nesterov step parameter
for k = 1:max_iter
  lambda_new = max(y - (Q * y - p)/nQ2, 0);  % Projected gradient step
  t_new = (1 + sqrt(1 + 4*t^2)) / 2; % Nesterov update
  y     = lambda_new + ((t - 1)/t_new) * (lambda_new - lambda);
  lambda = lambda_new;
  t = t_new;
d2(k) = 0.5*norm(A'*lambda)^2 - lambda'*p;
end
```

```
%% Figure
subplot(121)
plot(d,'LineWidth',3),hold on,
plot(d2,'LineWidth',3),
legend('PGD','Fast PGD');grid on,axis tight,
xlabel('iteration','Interpreter','latex')
ylabel('$d(\lambda)$','Interpreter','latex')
set(gca,'fontsize',24)

subplot(122)
dmin = min([min(d) min(d2)]);
semilogy(d-dmin,'LineWidth',3),hold on,
semilogy(d2-dmin,'LineWidth',3)
legend('PGD','Fast PGD');grid on,axis tight
xlabel('iteration','Interpreter','latex')
ylabel('$d(\lambda) - d_\textrm{min}$','Interpreter','latex')
set(gca,'fontsize',24)

%% Recover primal solution
x_star = u - A' * lambda
% find violated
find(A*x_star > b)
```

Quick experiment





Recall: after you get $\boldsymbol{\lambda}^*$, you get $\boldsymbol{x}^* = \boldsymbol{u} - \boldsymbol{A}^\top \boldsymbol{\lambda}^*$

When shall you use dual projection method

$\boldsymbol{x}$ has many variables, $\boldsymbol{A}$ has a few rows

$\implies$ low dimensional $\boldsymbol{\lambda}$

$\implies$ cheaper computation in the dual

$\implies$ you can use dual projection method

$\boldsymbol{x}$ has a few variables, $\boldsymbol{A}$ has many rows

$\implies$ high dimensional $\boldsymbol{\lambda}$

$\implies$ expensive computation in the dual

$\implies$ you should not use dual projection method

Contents

Why dual problem solves the original problem

- Original problem $(\mathcal{P}) : \underset{\boldsymbol{x}}{\operatorname{argmin}} \ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 \ \text{ s.t. } \ \boldsymbol{Ax} \leq \boldsymbol{b}$.

- The update is $\boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \dfrac{\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{\lambda}_k - (\boldsymbol{Au}-\boldsymbol{b})}{\|\boldsymbol{A}\boldsymbol{A}^\top\|_2}\right]_+$

  Not on $\boldsymbol{x}$

- Why the update on $\boldsymbol{\lambda}$ will solve $\mathcal{P}$????
  We explain it using Hoffman's Lemma

Hoffman's Lemma

- **Definition (Feasible set)** Let
$$F := \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}\} \quad \ne \varnothing$$
  - F is a set assumed to be non-empty
  - F is a polyhedron

- **Hoffman's Lemma**: There exists a constant $H_{\boldsymbol{A}} > 0$ such that for all $\boldsymbol{x} \in \mathbb{R}^n$,
$$\operatorname{dist}(\boldsymbol{x}, F) \le H \big\| [\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}]_+ \big\|_2.$$
  - $[\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}]_+$ is a nonlinear function, it returns the violation vector of $\boldsymbol{x}$
  - $\big\| [\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}]_+ \big\|_2$ is the "total amount of violation"
  - $H_{\boldsymbol{A}}$ depends on the matrix $\boldsymbol{A}$ only
  - Hoffman's Lemma is saying this number exists
  - The constant $H_{\boldsymbol{A}}$ is not explicitly known, but its existence is guaranteed.

- not discussed here
  - General version of the Hoffman's Lemma
  - How the Hoffman's Lemma is derived by Farkas' lemma

$H_{\boldsymbol{A}}$ represents geometric sensitivity

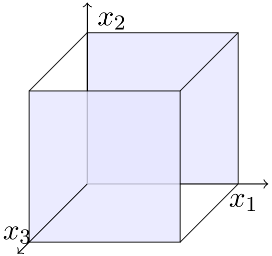
- The matrix $\boldsymbol{A}$ defines the orientation and structure of the polyhedron constraint set F.

- Hoffman's constant $H_{\boldsymbol{A}}$ captures how far an arbitrary point $\boldsymbol{x}$ can be from the set F, relative to its constraint violation $[\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}]_+$

- H captures the *geometric sensitivity* of the feasible set to constraint violations.

Example of A that has small $H_{\boldsymbol{A}}$

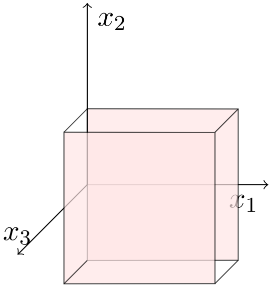
$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



- Well-spread (orthogonal) constraints: $x_1 \leq b_1, \quad x_2 \leq b_2, \quad x_3 \leq b_3$
- Geometry: Axis-aligned box
- Small violations imply small distance to feasible set $\Rightarrow$ small $H_{\boldsymbol{A}}$ (=1 here)

Example of A that has large $H_{\boldsymbol{A}}$

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0.001 & 0 \\ 1 & -0.001 & 0 \end{bmatrix}$$



- Constraints: All nearly aligned with x_1-axis
- Geometry: Thin slab
- Small constraint violations in $\boldsymbol{Ax} - \boldsymbol{b}$ can correspond to large distances to feasible set
- Poor geometry (too "flat") $\implies$ large $H_{\boldsymbol{A}}$ (proportion to $\kappa(\boldsymbol{A})$ here)

Hoffman's Lemma on the dual

$$\nabla d(\boldsymbol{\lambda}) = \boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{\lambda} - (\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b})$$

$$\begin{aligned}
\boldsymbol{x}_k &= \boldsymbol{u} - \boldsymbol{A}^\top\boldsymbol{\lambda}_k \\
\boldsymbol{A}\boldsymbol{x}_k &= \boldsymbol{A}\boldsymbol{u} - \boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{\lambda}_k \\
\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b} &= -\nabla d(\boldsymbol{\lambda}_k) \\
[\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}]_+ &= [-\nabla d(\boldsymbol{\lambda}_k)]_+ \\
[-\nabla d(\boldsymbol{\lambda}_k)]_+ &= [\nabla d(\boldsymbol{\lambda}_k)]_- \\
\operatorname{dist}(\boldsymbol{x}_k, F) &\leq H \big\| [\nabla d(\boldsymbol{\lambda}_k)]_- \big\|_2
\end{aligned}$$

- The norm of the **negative part of the dual gradient** bounds the primal feasibility violation:
  - Dual ProjGD reduces $\nabla d(\boldsymbol{\lambda}_k)$
  - By Hoffman, this drives $\boldsymbol{x}_k$ closer to feasibility
- Thus, dual projected gradient descent not only optimizes the dual but also **enforces primal feasibility**, with Hoffman's Lemma giving a quantitative guarantee.

Last page - summary

- Projection onto $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$
- $(\mathcal{P}) : \underset{\boldsymbol{x}}{\operatorname{argmin}} \ \dfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \ \text{ s.t. } \ \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}.$
- Lagrangian $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \dfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \rangle$
- Dual $d(\boldsymbol{\lambda}) = -\dfrac{1}{2}\|\boldsymbol{A}^\top \boldsymbol{\lambda}\|_2^2 + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b} \rangle$
- Dual problem $(\mathcal{D}) : \underset{\boldsymbol{\lambda} \geq \mathbf{0}}{\operatorname{argmin}} \ d(\boldsymbol{\lambda}) = \dfrac{1}{2}\|\boldsymbol{A}^\top \boldsymbol{\lambda}\|_2^2 - \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b} \rangle$
- Hoffman's Lemma $\operatorname{dist}\Big(\boldsymbol{x}, \ \big\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\big\}\Big) \leq H \Big\| [\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}]_+ \Big\|_2$

End of document