The Implicit Dual Newton Method: Final Report

Chieh Lin
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
chiehl1@cs.cmu.edu

Tzu-Ming Kuo
Language Technology Institute
Carnegie Mellon University
Pittsburgh, PA 15213
tzumingk@andrew.cmu.edu

Po-Wei Wang
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
poweiw@cs.cmu.edu

1 Introduction

Consider minimizing a general convex optimization problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad f(\boldsymbol{x}), \quad \text{subject to } g_i(\boldsymbol{x}) \leq 0,\ \forall i = 1 \ldots m. \tag{1}$$

To solve this problem, we can apply the primal-dual interior point method, which approximates the original objective $f(\boldsymbol{x})$ and contraints $g_i(\boldsymbol{x})$ by second-order Taylor expansions, examines the KKT condition of the approximation, and then solves the KKT system by Newton method. On the other hand, we know that a dual problem corresponding to the prime problem exist. Assuming zero duality gap, can we simply solve the primal-dual pair without approximation? Unfortunately, unlike quadratic program or semi-definite program, there are no analytic dual problems for most convex primal problems. This makes the optimization involving the dual problem harder. However, we can show that, by implicit function theorem, there is a link between the primal and dual derivatives (Dontchev and Rockafellar, 2009). We can actually derive the gradient and Hessian for the dual problem from the proximal functions of f and g_i, solving the dual optimization implicitly (Wang et al., 2016). This is called the implicit dual Newton method. In this project, we plan to apply the implicit dual Newton method to different optimization problems, and think about the convergence rate of the method when the proximal function is not solved exactly.

2 Implicit function theorem

Consider the dual of (1),

$$L(\boldsymbol{x}, \boldsymbol{\alpha}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \alpha_i g_i(\boldsymbol{x}). \tag{2}$$

The dual objective of (1) can be written as

$$D(\boldsymbol{\alpha}) = \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\alpha}). \tag{3}$$

While $D(\boldsymbol{\alpha})$ might not have an analytic form, minimizer $\boldsymbol{x}$ of (3) need to satisfy the following equality

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\alpha}) = \mathbf{0}. \tag{4}$$

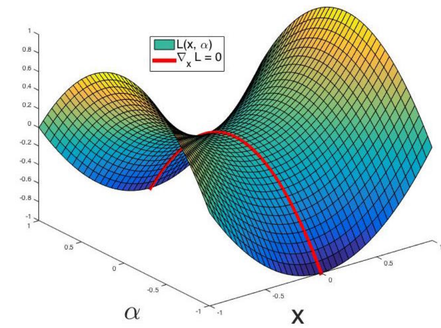


Figure 1: Illustration of the Lagrangian. The red curve corresponds to $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\alpha}) = \mathbf{0}$.

See Figure 1 for illustration of the curve $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\alpha}) = \mathbf{0}$. This means that $\boldsymbol{x}$ is actually a (general) function of $\boldsymbol{\alpha}$ in the dual problem. Thus, we can use the implicit function theorem to link the derivatives between primal and dual variables. That is,

$$\left(\frac{\partial}{\partial \boldsymbol{x}} \nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \alpha)\right)^{\top} d\boldsymbol{x} + \left(\frac{\partial}{\partial \boldsymbol{\alpha}} \nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\alpha})\right)^{\top} d\boldsymbol{\alpha} = \mathbf{0}. \tag{5}$$

This way, we can derive $\frac{d\boldsymbol{x}}{d\boldsymbol{\alpha}}$ and use the chain rule to find the derivatives of the dual objective (3). This method can be further generalize to other implicit convex functions, for example the point-wise maximum of convex functions, that have internal variables satisfying an equality like (4).

To demonstrate the power of the implicit dual method, we consider the following applications.

3 Application: Epigraph Projection

The epigraph projection problem, which projects the pair of point $\boldsymbol{u}$ and function values v to the epigraph set of the function epi$f(\cdot)$, is a key ingredient in the operator splitting algorithms. It can be cast to the following optimization problem,

$$\min_{\boldsymbol{x}, t} \|(\boldsymbol{x}, t) - (\boldsymbol{u}, v)\|^2/2, \quad \text{s.t. } f(\boldsymbol{x}) \leq t. \tag{epi}$$

In particular, we can show that the problem (epi) is equivalent to the following dual problem

$$\max_{\lambda \geq 0} \min_{\boldsymbol{x}, t} \|(\boldsymbol{x}, t) - (\boldsymbol{u}, v)\|^2/2 + \lambda(f(\boldsymbol{x}) - t), \tag{6}$$

which has only one variable λ. However, because $f(\cdot)$ can be arbitrary convex function, the dual of the epigraph projection problem cannot be expressed in an analytic form in general. For example, consider the epigraph projection for the entropy function $f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i \log(x_i)$, which is defined on the probability simplex

$$\boldsymbol{\Delta}_n = \{\boldsymbol{x} \geq \mathbf{0},\ \mathbf{1}^{\top} \boldsymbol{x} = 1,\ \boldsymbol{x} \in \mathbb{R}^n\}. \tag{7}$$

See Figure 2 for illustration. The optimization problem for the epigraph projection is

$$\min_{\boldsymbol{x}, t} \|(\boldsymbol{x}, t) - (\boldsymbol{u}, v)\|^2/2, \quad \text{s.t.} \quad \boldsymbol{x}^T \log(\boldsymbol{x}) \leq t, \quad x_i \geq 0 \quad \forall i, \quad 1^T \boldsymbol{x} = 1 \tag{8}$$

Deriving the dual epigraph projection problem of the entropy function $f(\boldsymbol{x})$ requires solving an log-linear equation similar to $\log(z) = z$, which does not has an analytic solution. Now we show that the implicit function theorem can be applied to solve the epigraph projection problem.

3.1 Implicit Dual Newton Method

If we fix the dual variable λ, we can observe that the solution of $\boldsymbol{x}$ is simply a proximal operator,

$$\boldsymbol{x} = \text{prox}_{\lambda}(\boldsymbol{u}) = \operatorname*{argmin}_{\boldsymbol{x}} \|\boldsymbol{x} - \boldsymbol{u}\|^2/2 + \lambda f(\boldsymbol{x}). \tag{9}$$

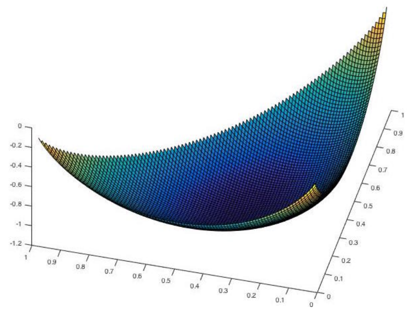


Figure 2: Illustration of the entropy function on the probability simplex Δ_3.

Thus, $\boldsymbol{x}$ can be considered to be a "function" of the dual variable λ, and we can applied the implicit function theorem to obtain the derivatives between the primal and dual variables,

$$\frac{dx}{d\lambda} = -(I + \lambda \nabla^2 f(\boldsymbol{x}))^{-1} \nabla f(\boldsymbol{x}). \tag{10}$$

This way, we can derive the gradient and Hessian of the dual function implicitly as following.

$$\frac{dD(\lambda)}{d\lambda} = \frac{dL(\boldsymbol{x}(\lambda), t(\lambda), \lambda))}{d\lambda} = f(\boldsymbol{x}) - v - \lambda \tag{11}$$

$$\frac{d^2 D(\lambda)}{d\lambda^2} = \frac{df}{d\boldsymbol{x}}^\top (I + \lambda \nabla^2 f(\boldsymbol{x}))^{-1} \frac{df}{d\boldsymbol{x}} - 1. \tag{12}$$

With the derivatives, we can apply the Newton method on the dual epigraph projection problem. See Algorithm 1 for details. The key for implicit dual Newton method to work in this case is that we usually have an efficient proximal operator for the inner problem of (6). In our experiment, we take

$$f(x) = \sum_{i=1}^{n} x_i \log x_i, \quad \text{where } \boldsymbol{x} \in \text{ the probability simplex } \Delta_n, \tag{13}$$

and the inner problem of (6) is equivalent to solving

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \quad \|\boldsymbol{x} - \boldsymbol{u}\|^2/2 + \lambda \sum_{i=1}^{n} x_i \log x_i, \tag{14}$$

$$\text{s.t.} \quad \mathbf{1}'\boldsymbol{x} = 1, \quad \boldsymbol{x} \geq 0. \tag{15}$$

The proximal problem is easy to solve in parallel by Newton method if we do not have the constraint $\mathbf{1}'\boldsymbol{x} = 1$. To solve this, we will apply the implicit function theorem again. The Lagrangian for the proximal problem is

$$L(x, \theta, \boldsymbol{\gamma}) = \|\boldsymbol{x} - \boldsymbol{u}\|^2/2 + \lambda \sum_{i=1}^{n} x_i \log x_i - \theta(\mathbf{1}'\boldsymbol{x} - 1) - \boldsymbol{\gamma}'\boldsymbol{x}. \tag{16}$$

By inspection, $x_i = 0$ is never reached, and all we need is to maintain the optimal condition:

$$\nabla_{\boldsymbol{x}} L = \boldsymbol{x} - \boldsymbol{u} + \lambda(\mathbf{1} + \log \boldsymbol{x}) - \theta \mathbf{1} = \mathbf{0}. \tag{17}$$

This means that $\boldsymbol{x}$ and θ are implicit function, and

$$\frac{dx_i}{d\theta} = \frac{x_i}{x_i + \lambda}. \tag{18}$$

To make $\boldsymbol{x}$ stay on $\mathbf{1}'\boldsymbol{x} = 1$, we solve

$$\min_{\theta} \ \|\mathbf{1}'\boldsymbol{x}(\theta) - 1\|^2 \tag{19}$$

by chain rule and do Newton method on (14) without considering the equality constraint iteratively. The proximal runs in $O(n)$ and usually converges in 5 iterations to 10^{-10}. The resulting implicit dual Newton method usually converges in 10 iterations to 10^{-10}.

Algorithm 1 Implicit dual newton method for epigraph projection

$\lambda := 1$
while not yet converged **do**
  $x := \text{prox}_{\lambda f}(v)$;
  $\frac{dD(\lambda)}{d\lambda} := f(x) - \lambda - s$;
  $\frac{d^2 D(\lambda)}{d\lambda^2} := -\nabla f(x)^\top (I + \lambda \nabla^2 f(x))^{-1} \nabla f(x) - 1$;
  if $|\frac{dD(\lambda)}{d\lambda}| \leq \epsilon$ **then** break;
  $\lambda := \max(0, \lambda - \frac{dD(\lambda)}{d\lambda} / \frac{d^2 D(\lambda)}{d\lambda^2})$;
end while
return $(\text{prox}_\lambda\, f(v), s + \lambda)$.

3.2 Barrier Method

Barrier method is another well-known way to solve (1). The inequality constraints are put into log barrier functions so that the problem is transformed to the following unconstrained convex optimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}} \quad \lambda f(\boldsymbol{x}) - \sum_{i=1}^{m} \log(g_i(\boldsymbol{x})). \tag{20}$$

The λ is initially set to be a small number, and we iteratively solve (20) by Newton method and increase λ by a factor μ until $f(\boldsymbol{x})$ converges.

Consider the epigraph of the entropy function $f(x) = \sum_{i=1}^{n} x_i \log(x_i)$ mentioned above.

The problem to be solved in each iteration of barrier method is

$$\begin{aligned} \min_{\boldsymbol{x}, t} \quad & g(\boldsymbol{x}, t) \equiv \lambda \|(\boldsymbol{x}, t) - (\boldsymbol{u}, v)\|^2/2 - \log(t - \boldsymbol{x}^T \log(\boldsymbol{x})) - \sum_{i=1}^{n} \log(x_i) \\ \text{s.t.} \quad & 1^T \boldsymbol{x} = 1. \end{aligned} \tag{21}$$

The gradient and Hessian of g are, respectively,

$$\nabla g(\boldsymbol{x}, t) = \begin{pmatrix} \lambda(\boldsymbol{x} - \boldsymbol{u}) + \frac{1}{t - \boldsymbol{x}^T \log(\boldsymbol{x})} (\log(\boldsymbol{x}) + 1) \\ \lambda(t - v) - \frac{1}{t - \boldsymbol{x}^T \log(\boldsymbol{x})} \end{pmatrix}, \tag{22}$$

$$\nabla^2 g(\boldsymbol{x}, t) = \lambda I_{n+1} + \frac{1}{(t - \boldsymbol{x}^T \log(\boldsymbol{x}))^2} \boldsymbol{a}\boldsymbol{a}^T + \frac{1}{t - \boldsymbol{x}^T \log(\boldsymbol{x})} \text{diag}(\boldsymbol{b}), \tag{23}$$

$$\tag{24}$$

where $\boldsymbol{a} = (\log(x_1) + 1, \ldots, \log(x_n) + 1, -1)$ and $\boldsymbol{b} = (\frac{1}{x_1}, \ldots, \frac{1}{x_n}, 0)$.

We solve (21) by Newton method, where Newton direction $\boldsymbol{v}$ is obtained by solving the following linear system

$$\begin{pmatrix} \nabla^2 g(\boldsymbol{x}, t) & 1 \\ 1^T & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{v} \\ w \end{pmatrix} = -\begin{pmatrix} \nabla g(\boldsymbol{x}, t) \\ 1^T \boldsymbol{x} - 1 \end{pmatrix}. \tag{25}$$

3.3 Primal-Dual Interior Point Method

Like barrier method, primal-dual interior point method (Wright, 1997) can also be applied to solve the epigraph projection problem. It is usually more efficient than barrier method. In primal-dual interior point method we derive the Lagrangian of (epi) as follows

$$L(\boldsymbol{x}, t, \lambda, \boldsymbol{\gamma}, \theta) = \frac{1}{2}\|(\boldsymbol{x} - \boldsymbol{u})\|^2 + \frac{1}{2}(t - v)^2 + \lambda(\boldsymbol{x}^T \log(\boldsymbol{x}) - t) + \theta(1^T \boldsymbol{x} - 1) - \boldsymbol{\gamma}^T \boldsymbol{x} \tag{26}$$

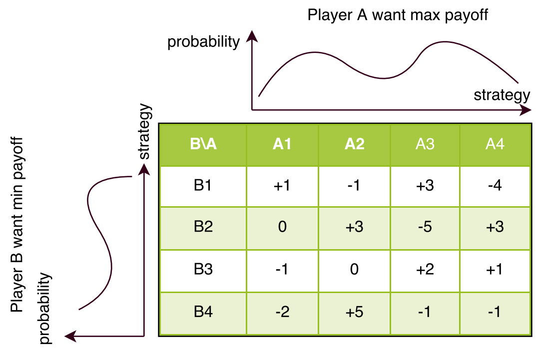


Figure 3: Illustration of a 2-player matrix game. Player A controls strategy $\boldsymbol{u}$ and want the cost to be minimized. Player B controls strategy $\boldsymbol{v}$ and want the cost to be maximized.

Where λ, θ, and $\boldsymbol{\gamma}$ are the dual variables for the constraints $\boldsymbol{x}^T \log(\boldsymbol{x}) \leq t$, $1^T \boldsymbol{x} = 1$, and $\boldsymbol{x} \geq 0$ respectively. We can write the KKT conditions for both primal and dual problems as follows.

$$\nabla_{\boldsymbol{x}} L = (\boldsymbol{x} - \boldsymbol{u}) + \lambda(\log(x) + 1) + \theta 1 - \boldsymbol{\gamma} \tag{27}$$

$$\nabla_t L = (t - v) - \lambda \tag{28}$$

$$\lambda \geq 0, \quad \boldsymbol{\gamma} \geq 0, \quad \boldsymbol{x} \geq 0, \quad 1^T \boldsymbol{x} = 1 \tag{29}$$

$$\lambda(t - \boldsymbol{x}^T \log(\boldsymbol{x})) = \tau, \quad \gamma_i x_i = \tau \quad \forall i \tag{30}$$

To satisfy all the KKT conditions, we need to make the residual $\boldsymbol{r}(\boldsymbol{x}, t, \lambda, \boldsymbol{\gamma}, \theta)$ to be zero, where

$$\boldsymbol{r}(\boldsymbol{x}, t, \lambda, \boldsymbol{\gamma}, \theta) = \begin{pmatrix} r_{\text{dual},\boldsymbol{x}} \\ r_{\text{dual},t} \\ r_{\text{cent},\lambda} \\ r_{\text{cent},\boldsymbol{\gamma}} \\ r_{\text{prime},\theta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x} - \boldsymbol{u} + \lambda \log(\boldsymbol{x}) + \lambda 1 + \theta 1 - \boldsymbol{\gamma} \\ (t - v) - \lambda \\ \tau - \lambda(t - \boldsymbol{x}^\top \log(\boldsymbol{x})) \\ (\tau - \gamma_i x_i)_i \\ 1^T \boldsymbol{x} - 1 \end{pmatrix} \tag{31}$$

To zero out the residual, we take the first-order Taylor's expansion

$$r(z + \Delta z) \approx r(z) + Dr(z)\Delta z = 0 \text{ , where } z = (\boldsymbol{x}, t, \lambda, \boldsymbol{\gamma}, \theta). \tag{32}$$

We use Newton's method with backtracking to solve Δz for our update in each iteration.

4 Application: Convex-Concave Game

To demonstrate that the implicit dual Newton method can also be generalized beyond convex optimization, we consider the convex-concave matrix game, which is important in game theory. Let $A \in \mathbb{R}^{m \times n}$ be the payoff matrix of a two-player game. In this game, one player controlling the probability vector $\boldsymbol{u} \in \boldsymbol{\Delta}_m$ wishes to minimize the expected payoff, and the other player controlling $\boldsymbol{v} \in \boldsymbol{\Delta}_n$ wishes to maximize the expected payoff. See Figure 3 for illustration. The matrix game can be formalized as the following minimax optimization problem.

$$\min_{\boldsymbol{u} \in \boldsymbol{\Delta}_m} \max_{\boldsymbol{v} \in \boldsymbol{\Delta}_n} \boldsymbol{u}^\top A \boldsymbol{v}. \tag{33}$$

To make the problem simpler, we add the entropy barrier and to create a more well-conditioned problem

$$\min_{\boldsymbol{u} \in \boldsymbol{\Delta}_m} \max_{\boldsymbol{v} \in \boldsymbol{\Delta}_n} \boldsymbol{u}^\top A \boldsymbol{v} - \mu \left(\log m + \sum_i v_i \log(v_i) \right). \tag{34}$$

We can observe that the above problem is convex in $\boldsymbol{u}$ and concave in $\boldsymbol{v}$. Further, the addition term in the formulation (34) is indeed a entropy-barrier of the original problem (33). Because the convex-concave matrix game is not simply a convex optimization problem, the usual tools in convex optimization cannot be applied. However, we will show that the implicit dual Newton method still works on this problem.

4.1 Implicit Dual Proximal Method

Thanks to the entropy barrier function, we can see that there is a closed-form solution for $\boldsymbol{v}$ given a fixed $\boldsymbol{u}$ by inspecting:

$$v_i = \frac{\exp(z_i)}{\sum_i \exp(z_i)}, \quad \text{where } \boldsymbol{z} = A'\boldsymbol{u}/\mu. \tag{35}$$

This way, $\boldsymbol{v}$ in the minimax problem can be regarded as a function of $\boldsymbol{u}$, and we can apply the implicit function theorem to obtain the relation between the $\boldsymbol{u}$ and $\boldsymbol{v}$,

$$\frac{d\boldsymbol{v}}{d\boldsymbol{u}} = \frac{1}{\mu} A \operatorname{diag}\left(\frac{\exp(z_i)}{\sum_i \exp(z_i)}\right). \tag{36}$$

Let the objective function be f. The gradient of $\boldsymbol{u}$ with regard to the optimal $\boldsymbol{v}(\boldsymbol{u})$ is

$$\nabla_{\boldsymbol{u}} \max_{\boldsymbol{v} \in \Delta_n} f(\boldsymbol{u}, \boldsymbol{v}) = \frac{df}{d\boldsymbol{u}}(\boldsymbol{u}, \boldsymbol{v}(\boldsymbol{u})) + \frac{d\boldsymbol{u}}{d\boldsymbol{v}} \frac{df}{d\boldsymbol{v}}(\boldsymbol{u}, \boldsymbol{v}(\boldsymbol{u})) \tag{37}$$

$$= A\boldsymbol{v} + \frac{1}{\mu} A \operatorname{diag}\left(\frac{\exp(z_i)}{\sum_i \exp(z_i)}\right) (A'\boldsymbol{u} - \mu(\boldsymbol{1} + \log \boldsymbol{v})). \tag{38}$$

Then we can apply projected gradient method on $\boldsymbol{u}$.

4.2 Primal-Dual Interior Point Method

Primal-dual interior method can also be applied to convex-concave problems. We can derive the Lagrangian and write the KKT conditions for both the primal and the dual problems. The Lagrangian for (34) is

$$L(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{x}, \theta, \lambda) = \boldsymbol{u}^\top A\boldsymbol{v} - \mu \log m - \mu \boldsymbol{v}^T \log(\boldsymbol{v}) + \boldsymbol{y}^T \boldsymbol{v} - \boldsymbol{x}^T \boldsymbol{u} + \theta(1^T \boldsymbol{v} - 1) - \lambda(1^T \boldsymbol{u} - 1) \tag{39}$$

Where $\boldsymbol{y}$, $\boldsymbol{x}$, θ, and λ are the dual variables for the constraints $\boldsymbol{v} \geq 0$, $\boldsymbol{u} \geq 0$, $1^T\boldsymbol{v} = 1$, and $1^T\boldsymbol{u} = 1$ respectively. We write the KKT conditions for both primal and dual problems as follows.

$$\nabla_{\boldsymbol{v}} L = A^T \boldsymbol{u} - \mu(\log(\boldsymbol{v}) + 1) + \boldsymbol{y} + \theta \tag{40}$$

$$\nabla_{\boldsymbol{u}} L = A\boldsymbol{v} - \boldsymbol{x} - \lambda \tag{41}$$

$$\boldsymbol{x} \geq 0, \quad \boldsymbol{y} \geq 0, \quad \boldsymbol{v} \geq 0, \quad \boldsymbol{u} \geq 0, \quad 1^T\boldsymbol{v} = 1, \quad 1^T\boldsymbol{u} = 1 \tag{42}$$

$$\boldsymbol{y}_i \boldsymbol{v}_i = \tau, \quad \boldsymbol{x}_i \boldsymbol{u}_i = \tau \quad \forall i \tag{43}$$

Similar to section 3.3, we derive the following residual with regard to the primal and dual KKT conditions:

$$\begin{pmatrix} r_{\text{dual},\boldsymbol{v}} \\ r_{\text{dual},\boldsymbol{u}} \\ r_{\text{cent},\boldsymbol{y}} \\ r_{\text{cent},\boldsymbol{x}} \\ r_{\text{prim},\theta} \\ r_{\text{prim},\lambda} \end{pmatrix} = \begin{pmatrix} A^T\boldsymbol{u} - \mu(\log(\boldsymbol{v}) + 1) + \boldsymbol{y} + \theta \\ A\boldsymbol{v} - \boldsymbol{x} - \lambda \\ \tau - (y_i v_i)_i \\ \tau - (u_i x_i)_i \\ 1^T\boldsymbol{v} - 1 \\ 1^T\boldsymbol{u} - 1 \end{pmatrix}. \tag{44}$$

Then we can derive the first-order Taylor expansion to annihilate the residual. We apply the Newton's method with backtracking to solve this problem.

4.3 Accelerated Proximal Gradient Method

In addition to primal-dual interior point method, we can also apply the accelerated proximal gradient method (Tseng, 2008) to solve the convex-concave matrix game. The method extends Nemirovski's prox-type method in (Nemirovski, 2004) to solve convex-concave problem as follows.

Algorithm 2 Accelerated Projected Gradient Method

Assume l_f is L-Lipschitz continous

while *not converges* **do**

Choose $\gamma_k > 0$, $\bar{\boldsymbol{u}} \in \mathbb{R}^m$, and $\bar{\boldsymbol{v}} \in \mathbb{R}^n$ satisfying

$$\min_{\boldsymbol{u}\in\mathbb{R}^m, \boldsymbol{v}\in\mathbb{R}^n} \left\{ l_f(\boldsymbol{u}, \boldsymbol{v}; \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) + \frac{L}{\gamma_k} D(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{u}_k, \boldsymbol{v}_k) \right\} \geq l_f(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}; \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) \tag{45}$$

Let $\boldsymbol{u}_{k+1}, \boldsymbol{v}_{k+1}$ attain the minimum in (45). $k \leftarrow k+1$.

end

Note that

$$l_F(\boldsymbol{u}, \boldsymbol{v}; \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}) = \begin{bmatrix} \boldsymbol{u} & \boldsymbol{v} \end{bmatrix} \begin{bmatrix} A^T \bar{\boldsymbol{v}} \\ -A\bar{\boldsymbol{u}} \end{bmatrix} + \delta_{\Delta_m}(\boldsymbol{u}) + \delta_{\Delta_n}(\boldsymbol{v}),$$

$$\begin{aligned} D(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{u}_k, \boldsymbol{v}_k) = &\sum_{i=1}^m \Big(u_i \log u_i - u_{ki} \log u_{ki} - (\log u_{ki} + 1)(u_i - u_{ki}) \Big) \\ &+ \sum_{i=1}^n \Big(v_i \log v_i - v_{ki} \log v_{ki} - (\log v_{ki} + 1)(v_i - v_{ki}) \Big). \end{aligned}$$

To choose appropriate $\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}$, and γ_k, we simply let $\gamma_k = 1$ and $\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}$ to be the optimal solution of the following problem, which admits a close form solution,

$$\min_{\boldsymbol{u}\in\mathbb{R}^m, \boldsymbol{v}\in\mathbb{R}^n} \left\{ l_f(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{u}_k, \boldsymbol{v}_k) + \frac{L}{\gamma_k} D(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{u}_k, \boldsymbol{v}_k) \right\}. \tag{46}$$

Replacing the probability simplex indicator functions $\delta_{\Delta_m}(\boldsymbol{u})$ and $\delta_{\Delta_n}(\boldsymbol{v})$ to equality and inequality constraints, we can derive the KKT conditions for (46), which leads to the optimal solution

$$\begin{aligned} \boldsymbol{u}^* &\propto \exp\left(\frac{\gamma_k}{L} A^T \boldsymbol{v}_k + \log(\boldsymbol{u}_k) \right), \\ \boldsymbol{v}^* &\propto \exp\left(-\frac{\gamma_k}{L} A \boldsymbol{u}_k + \log(\boldsymbol{v}_k) \right). \end{aligned}$$

For the cases where Lipschitz constant L is unknown beforehand, we can have a initial guess of L and check if the solution of (46) satisfies (45). If not, we keep increasing L by a factor of 2 until the solution of (46) satisfies (45). In fact, the accelerated proximal gradient on matrix game is equivalent to Korpelevich's extragradient method (Korpelevich (1976)) with a Bregman function.

5 Experiments

In the experiment, we compare the benchmark method on random-generated dataset. For the epigraph projection problem, we generate the $\boldsymbol{u} \in \mathbb{R}^{5000}$ vector from a normalized uniform distribution, and set the value of $t = 4f(\boldsymbol{u})$ to make the initial $(\boldsymbol{u}, v)$ pair infeasible to the epigraph set (so that the problem is non-trivial). For the matrix game problem, we generate the payoff matrix $A \in \mathbb{R}^{500\times500}$ from uniform distribution of $[-1, 1]$ plus some rank-one Bernoulli tensor. For epigraph projection algorithm, we plot the function difference to the optimal value versus the iterations. For the convex-concave game problem, we plot the absolute value to the optimal solution versus the iterations. Note that the convex-concave game problem is a minimax problem. That is, the objective may not be strictly decreasing.

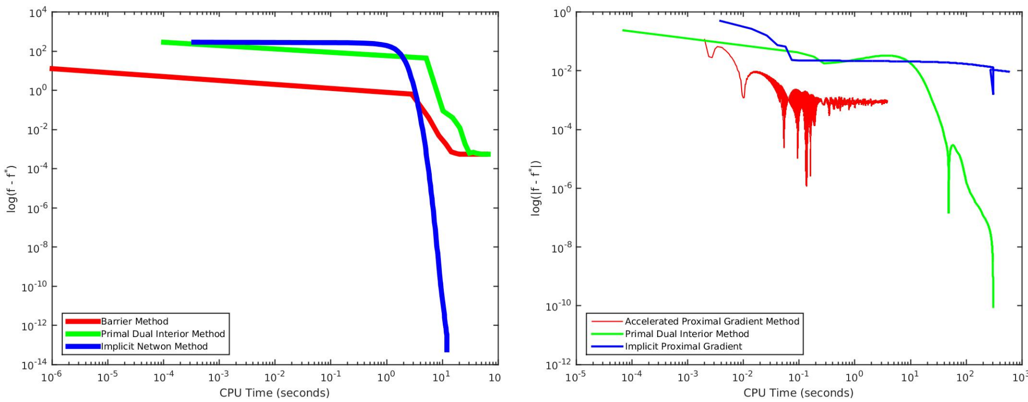
For the epigraph projection, we can see that the implicit Newton method did converges faster than the other method, possibly because we are able to get the proximal is cheap. For convex-concave game, the implicit proximal gradient method converges slower than other second-order methods, because it is only has first-order convergence.

References

Dontchev, A. L. and Rockafellar, R. T. (2009). Implicit functions and solution mappings.

(a) Epigraph Projection

(b) Convex-concave Game



Korpelevich, G. (1976). The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756.

Nemirovski, A. (2004). Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251.

Tseng, P. (2008). On accelerated proximal gradient methods for convex-concave optimization. submitted to siam j. *J. Optim*.

Wang, P.-W., Wytock, M., and Kolter, J. Z. (2016). Epigraph projections for fast general convex programming. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2868–2877.

Wright, S. J. (1997). *Primal-dual interior-point methods*. Siam.