Accelerated Primal-Dual Methods for Convex-Strongly-Concave Saddle Point Problems

Mohammad Khalafi, and Digvijay Boob

Department of Operations Research and Engineering Management, Southern Methodist University

Saddle Point Problems

We can convert many classical convex optimization problems with smooth or nonsmooth objective functions into a saddle point problem as (1).

$$\mathcal{L}(x,y) := \min_{x\in X}\max_{y\in Y} f(x) + \phi(x,y) - g(y). \quad (1)$$

Our convergence rate measure at $\bar{z} = (\bar{x},\bar{y})$ is:

$$Gap(\bar{z}) = \max_{z\in X\times Y}\{Q(\bar{z},z) := \mathcal{L}(\bar{x},y) - \mathcal{L}(x,\bar{y})\}.$$

Also, $\phi(\cdot,y)$ is L_{xx}-smooth, $\phi(x,\cdot)$ is L_{yy}-smooth and ϕ is L_{xy}-smooth, if the followings hold for all $x,x'\in X,\ y,y'\in Y$ respectively:

$$\|\nabla_x\phi(x',y) - \nabla_x\phi(x,y)\| \le L_{xx}\|x'-x\|,$$
$$\|\nabla_y\phi(x,y') - \nabla_y\phi(x,y)\| \le L_{yy}\|y'-y\|,$$
$$\|\nabla_y\phi(x',y) - \nabla_y\phi(x,y)\| \le L_{xy}\|x'-x\|.$$

Motivation of Study

In many problems, the following function is a nonsmooth function which is hard to optimize.

$$P(x) : f(x) + \max_{y\in Y}\phi(x,y). \quad (2)$$

One way to smoothen this function is to use Nesterov's smoothing technique. This technique involves subtracting a strongly convex regularizing function ([1]). This regularizing function is g in our model (see equation (1)). Therefore, the corresponding SPP is a **convex-strongly-concave** problem where g is μ_g-strongly convex. Such setting will be useful in ML problems with a complex constraint set. Furthermore, in many settings, we assume that $f(x)$ is an easy function to evaluate. This might not be true in many cases. Hence, linearization of f might be a good approach to handle this problem. In this context, one popular approach is using a linearized primal-dual method (LPD) ([2]). In this study, we investigate an LPD method for a convex-strongly-concave SPP.

Linearized Primal-Dual method: An important observation

Consider problem (1) with $\phi = \langle Ax, y\rangle$. Thus, the corresponding LPD can be shown as Algorithm 1

Algorithm 1 Linearized PD (LPD) method

Initialize $\tilde{x}_1 = x_1 \in X,\ y_1 \in Y$
for $t = 1,\dots,K$ **do**
$y_{t+1} \leftarrow \arg\min_{y\in Y}\langle -A\tilde{x}_t, y\rangle + g(y) + \frac{1}{2\tau_t}\|y - y_t\|^2$
$x_{t+1} \leftarrow \arg\min_{x\in X}\langle \nabla f(x_t) + A^\top y_{t+1}, x\rangle + \frac{1}{2\eta_t}\|x - x_t\|^2$
$\tilde{x}_{t+1} \leftarrow x_{t+1} + \theta_t(x_{t+1} - x_t)$
end for
return $\bar{x}_{K+1} = \frac{\sum_{t=1}^K \gamma_{t+1}x_{t+1}}{\sum_{t=1}^K\gamma_{t+1}}, \bar{y}_{K+1} = \frac{\sum_{t=1}^K\gamma_{t+1}y_{t+1}}{\sum_{t=1}^K\gamma_{t+1}}$

Convergence analysis of LPD

For a μ_f-strongly-convex-concave bilinear SPP, LPD has the optimal convergence rate of $\mathcal{O}(\frac{L_f+\|A\|^2}{K^2})$, and for a μ_g-strongly-concave-convex bilinear SPP, it has convergence rate of $\mathcal{O}(\frac{L_f}{K} + \frac{\|A\|^2}{K^2})$ where f is L_f-smooth.

Observation: Strong concavity can not handle the errors caused by the linearization of f.

Accelerated LPD (ALPD) for a general $\phi(x,y)$: A remedy

Algorithm 2 Accelerated Linearized PD (ALPD) method

Initialize $\bar{x}_1 = x_0 = x_1 \in X, \bar{y}_1 = y_0 = y_1 \in Y$
for $t = 1,\dots,K$ **do**
$\underline{x}_t \leftarrow (1-\beta_t^{-1})\bar{x}_t + \beta_t^{-1}x_t$
$v_t \leftarrow (1+\theta_t)\nabla_y\phi(x_t,y_t) - \theta_t\nabla_y\phi(x_{t-1},y_{t-1})$
$y_{t+1} \leftarrow \arg\min_{y\in Y}\langle -v_t + \nabla g(y_t), y\rangle + \frac{1}{2\tau_t}\|y - y_t\|^2$
$x_{t+1} \leftarrow \arg\min_{x\in X}\langle \nabla f(\underline{x}_t) + \nabla_x\phi(x_t,y_{t+1}), x\rangle + \frac{1}{2\eta_t}\|x - x_t\|^2$
$\bar{x}_{t+1} = (1-\beta_t^{-1})\bar{x}_t + \beta_t^{-1}x_{t+1}$
$\bar{y}_{t+1} = (1-\beta_t^{-1})\bar{y}_t + \beta_t^{-1}y_{t+1}$
end for
return $\bar{x}_{K+1}, \bar{y}_{K+1}$

Convergence rates of ALPD for semi-linear and nonlinear coupling

- *Case 1: Semi-linear ϕ with $L_{xx} = 0$:*
$$\max_{z\in X\times Y}\{Q(\bar{z}_{K+1})\} = \mathcal{O}(\tfrac{L_f+L_{yy}}{K^2} + \tfrac{L_{xy}^2}{\mu_g K^2})$$
- *Case 2: nonlinear ϕ with $L_{xx} > 0$:*
$$\max_{z\in X\times Y}\{Q(\bar{z}_{K+1})\} = \mathcal{O}(\tfrac{L_f+L_{yy}}{K^2} + \tfrac{L_{xy}^2}{\mu_g K^2} + \tfrac{L_{xx}}{K})$$

Inexact ALPD

As we see, ALPD has $\mathcal{O}(\frac{L_{xx}}{\epsilon})$ gradient complexity in $\nabla_x\phi$. We propose the following inexact ALPD to improve this gradient complexity. Algorithm 3, is a two-loop algorithm that solves a proximal problem using AGD in the inner loop while the outer loop follows a "conceptual" ALPD method.

Algorithm 3 Inexact ALPD Method

Initialize $\bar{x}_1 = x_0 = x_1 \in X, \bar{y}_1 = y_0 = y_1 \in Y$
for $t = 1,\dots,K$ **do**
$\underline{x}_t \leftarrow (1-\beta_t^{-1})\bar{x}_t + \beta_t^{-1}x_t$
$v_t \leftarrow (1+\theta_t)\nabla_y\phi(x_t,y_t) - \theta_t\nabla_y\phi(x_{t-1},y_{t-1})$
$y_{t+1} \leftarrow \arg\min_{y\in Y}\langle -v_t + \nabla g(y_t), y\rangle + \frac{1}{2\tau_t}\|y - y_t\|^2$
x_{t+1} is a δ_t-approximate solution of the problem:
$$\min_{x\in X}\langle \nabla f(\underline{x}_t), x\rangle + \phi(x, y_{t+1}) + \tfrac{1}{2\eta_t}\|x - x_t\|^2$$
$\bar{x}_{t+1} \leftarrow (1-\beta_t^{-1})\bar{x}_t + \beta_t^{-1}x_{t+1}$
$\bar{y}_{t+1} \leftarrow (1-\beta_t^{-1})\bar{y}_t + \beta_t^{-1}y_{t+1}$
end for
return $\bar{x}_{K+1}, \bar{y}_{K+1}$

Complexity analysis of inexact ALPD

Inexact ALPD requires $\mathcal{O}(\sqrt{\frac{L_f+L_{yy}}{\epsilon}})$ gradient evaluation of ∇f and $\nabla_y\phi$, and requires $\mathcal{O}(\frac{\sqrt{L_{xx}}}{\epsilon^{3/4}}\log(\frac{1}{\epsilon})) = \tilde{\mathcal{O}}(\frac{\sqrt{L_{xx}}}{\epsilon^{3/4}})$ gradient evaluation of $\nabla_x\phi$. Hence, the gradient complexity of $\nabla_x\phi$ improves significantly (c.f. $\mathcal{O}(\frac{L_{xx}}{\epsilon})$ gradient complexity in ALPD)

Numerical experiments

The smooth approximation of the nonsmooth penalty problem using Nesterov's smoothing technique is the following

$$\min_{x\in X}\max_{\|y\|_p\le 1}\{f(x) + \rho\langle y, Ax - b\rangle - \tfrac{\mu_g}{2}\|y\|^2\}, \quad (3)$$

where f is a quadratic function.

ALPD vs. LPD : Linear constraints

Note ALPD-prox-g is a variant of ALPD in which do not linearize g.

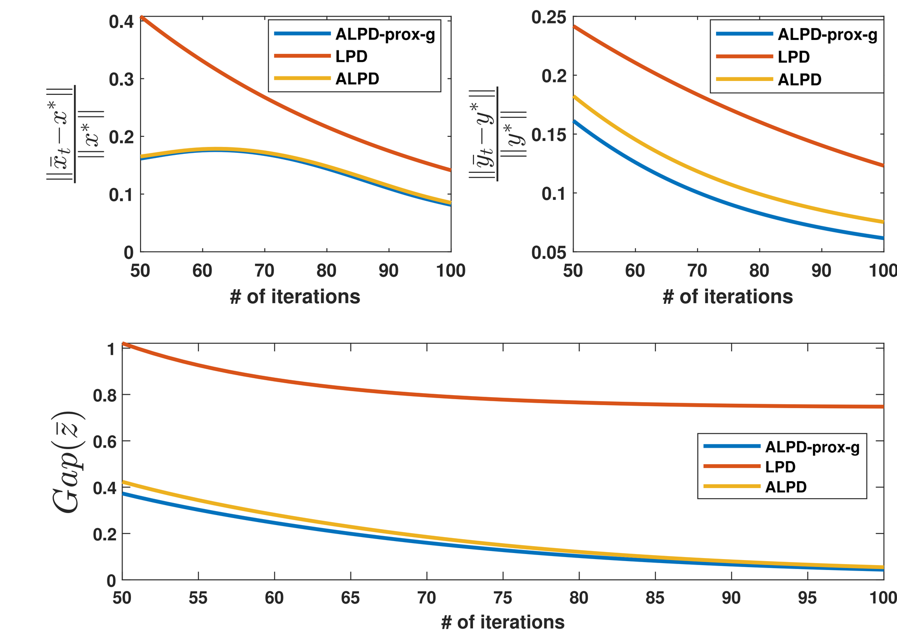

Figure 1: Comparison of the methods in terms of the mean errors in primal (top left), dual (top right) and Gap function (bottom) for 10 i.i.d. instances of (3) with $p = q = 2$.

ALPD vs. Inexact ALPD: quadratic constraints ($L_{xx} > 0$) .

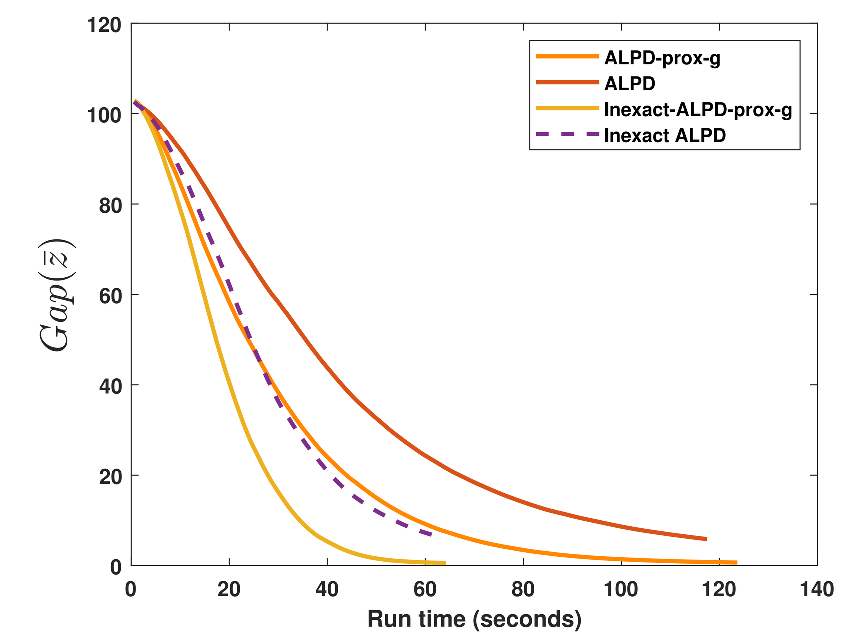

Figure 2: Comparison of the ALPD and inexact ALPD method and their prox-g variants using the Gap function vs run-time (seconds) plot for 10 i.i.d. instances.

LPD step-size policy: [2] vs. ours

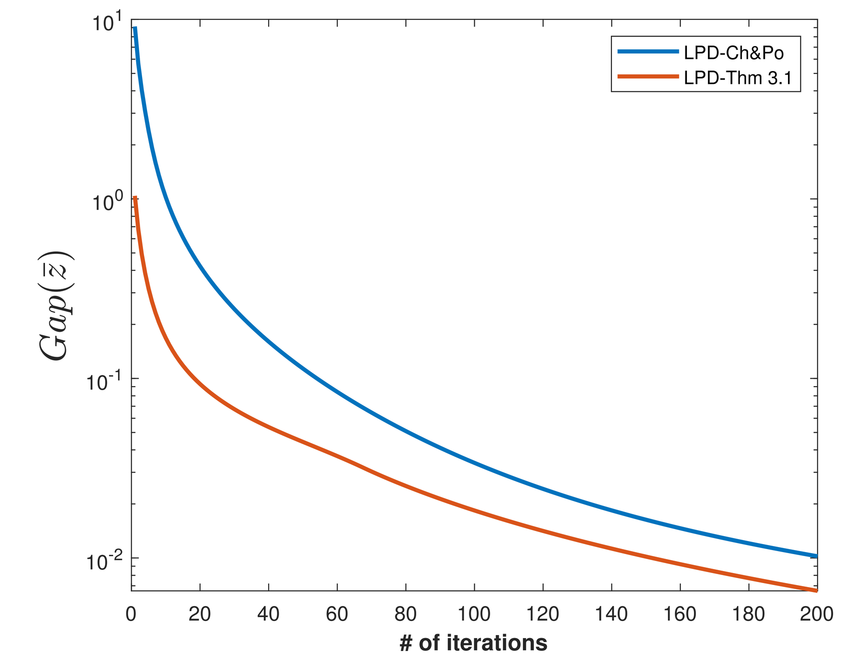

Figure 3: Comparison between the step-size policies of our work and [2] for 10 i.i.d. problem instances. Both policies start from the same initial point.

Summary

Algorithm	Coupling	Gradient Complexity	
		$\mu_f > 0$	$\mu_g > 0$
LPD	bilinear	$\mathcal{O}(1/\sqrt{\epsilon})$	$\mathcal{O}(L_f/\epsilon + \|A\|/\sqrt{\mu_g\epsilon})$
ALPD	semi-linear	NA	$\mathcal{O}(\sqrt{(L_f+L_{yy})/\epsilon} + L_{xy}/\sqrt{\mu_g\epsilon})$
ALPD	general	NA	$\mathcal{O}(\sqrt{(L_f+L_{yy})/\epsilon} + L_{xy}/\sqrt{\mu_g\epsilon} + L_{xx}/\epsilon)$
Inexact ALPD	general	NA	For $\nabla f, \nabla_y\phi:\ \mathcal{O}(\sqrt{(L_f+L_{yy})/\epsilon}$ For $\nabla_x\phi:\ \mathcal{O}(\frac{\sqrt{L_{xx}}}{\epsilon^{3/4}}\log(\frac{1}{\epsilon}))$

References

[1] Nesterov, Y. Smooth minimization of non-smooth functions. Math. Program. 103, 127–152 (2005)

[2] Chambolle, A., Pock, T. On the ergodic convergence rates of a first-order primal–dual algorithm. Math. Program. 159, 253–287 (2016)