Efficient quadratic penalization through the partial minimization technique

A.Y. Aravkin, D. Drusvyatskiy, and T. van Leeuwen

***Abstract*—Common computational problems, such as parameter estimation in dynamic models and PDE constrained optimization, require data fitting over a set of auxiliary parameters subject to physical constraints over an underlying state. Naive quadratically penalized formulations, commonly used in practice, suffer from inherent ill-conditioning. We show that surprisingly the so-called partial minimization technique regularizes the problem, making it well-conditioned. This viewpoint sheds new light on the penalty method for PDE constrained optimization and motivates robust extensions. In addition, we outline an inexact analysis, showing that the partial minimization subproblem can be solved very loosely in each iteration. We illustrate the theory and algorithms on boundary control, optimal transport, and parameter estimation for robust dynamic inference.**

I. Introduction

In this work, we consider a structured class of optimization problems having the form

$$\min_{y,u} \; f(y) + g(u) \qquad \text{subject to} \qquad A(u)y = q. \tag{1}$$

Here, $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex and smooth, $q \in \mathbb{R}^n$ is a fixed vector, and $A(\cdot)$ is a smoothly varying invertible matrix. For now, we make no assumptions on $g\colon \mathbb{R}^d \to \mathbb{R}$, though in practice, it is typically either a smooth or a 'simple' nonsmooth function. Optimization problems of this form often appear in PDE constrained optimization [6], [21], [23], Kalman filtering [2], [4], [15], boundary control problems [11], [19], and optimal transport [1], [13]. Typically, u encodes auxiliary variables while y encodes the state of the system; the constraint $A(u)y = q$ corresponds to a discretized PDE describing the physics.

Since the discretization $A(u)y = q$ is already inexact, it is appealing to relax it in the formulation. A seemingly naive relaxation approach is based on the quadratic penalty:

$$\min_{y,u} \; F(y,u) := f(y) + g(u) + \lambda \cdot \|A(u)y - q\|^2. \tag{2}$$

Here $\lambda > 0$ is a relaxation parameter for the equality constraints in (1), corresponding to relaxed physics. The classical quadratic penalty method in nonlinear programming proceeds by applying an iterative optimization algorithm to the unconstrained problem (2) until some termination criterion is satisfied, then increasing λ, and repeating the procedure with the previous iterate used as a warm start. For a detailed discussion, see e.g. [17, Section 17.1]. The authors of [22] observe that this strategy helps to avoid extraneous local minima, in contrast to the original formulation (1). From this consideration alone, the formulation (2) appears to be useful.

Conventional wisdom teaches us that the quadratic penalty technique is rarely appropriate. The difficulty is that one must allow λ to tend to infinity in order to force near-feasibility in the original problem (1); the residual error $\|A(u)y - q\|$ at an optimal pair (u, y) for the problem (2) is at best on the order of $\mathcal{O}(1/\lambda)$. Consequently, the maximal eigenvalue of the Hessian of the penalty term scales linearly with λ and the problems (2) become computationally difficult. Indeed, the maximal eigenvalue of the Hessian determines the behavior of numerical methods (gradient descent, quasi-Newton) far away from the solution – a regime in which most huge scale problems are solved. Figure 1 is a simple numerical illustration

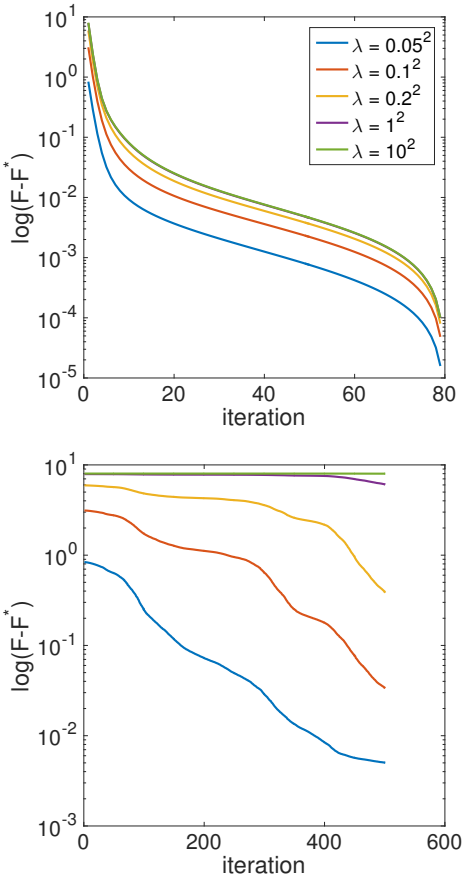


Fig. 1: Top panel shows $\log(F_k - F^*)$ of L-BFGS **with** partial minimization applied to (2) for the boundary control problem in Section III. Bottom panel shows $\log(F_k - F^*)$ of L-BFGS **without** partial minimization applied to (2), for the same values of λ. Both methods are initialized from the same starting points. Performance of L-BFGS without partial minimization degrades as λ increases, while performance of L-BFGS with partial minimization is insensitive to large λ.

of this inherent difficulty on a boundary control problem; see Section III for more details on the problem formulation. The bottom panel in the figure tracks progress of the objective function in (2) when an L-BFGS method is applied jointly in the variables (u, y). After 500 iterations of L-BFGS, the objective value significantly increases with increasing λ, while the actual minimal value of the objective function converges to that of (1), and so hardly changes. In other words, performance of the method scales poorly with λ, illustrating the ill-conditioning.

In this paper, we show that by using a simple *partial minimization* step, this complexity blow-up can be avoided entirely. The resulting algorithm is perfectly suited for many large-scale problems, where satisfying the constraint $A(u)y = q$ to high accuracy is not required (or even possible). The performance of the new method is shown in the top panel in Figure 1.

The strategy is straightforward: we rewrite (2) as

$$\min_u \ \widetilde{\varphi}(u) + g(u), \tag{3}$$

where the function $\widetilde{\varphi}(u)$ is defined implicitly by

$$\widetilde{\varphi}(u) = \min_y \left\{ f(y) + \lambda \cdot \|A(u)y - q\|^2 \right\}. \tag{4}$$

We will call $\widetilde{\varphi}(\cdot)$ the *reduced function*. Though this approach of *minimizing out* the variable y is often used (e.g. [7], [10], [22]), little theoretical justification for its superiority is known. In this work, we show that not only does partial minimization perform well numerically for the problem class (2), but is also theoretically grounded. We prove that surprisingly the Lipschitz constant of $\nabla\widetilde{\varphi}$ is bounded by a constant independent of λ. Therefore, iterative methods can be applied directly to the formulation (3). The top panel of Figure 1 is obtained by applying L-BFGS to (3). If $g(u)$ is nonsmooth, BFGS methods or prox-gradient methods can be applied.

The inner problem (4) can be solved efficiently since its condition number is nearly independent of λ. When f is a convex quadratic and $A(u)$ is sparse, one can apply sparse direct solvers or iterative methods such as LSQR [18]. More generally, when f is an arbitrary smooth convex function, one can apply first-order methods, which converge globally linearly with the rate governed by the condition number of the strongly convex objective in (4). Quasi-newton methods or variants of Newton's method are also available.

The outline of the paper is as follows. In Section II, we present complexity guarantees of the partial minimization technique. In Section III, we numerically illustrate the overall approach on boundary control and optimal transport problems, and on tuning an oscillator from very noisy measurements.

II. Theory

In this section, we show that the proposed framework is insensitive to the parameter λ. Throughout we assume that f and $A(\cdot)$ are C^2-smooth, f is convex, and $A(u)$ is invertible for every $u \in \mathbb{R}^n$.

To shorten the formulas, in this section, we will use the symbol A_u instead of $A(u)$ throughout. Setting the stage, define the function

$$\varphi(u, y) = f(y) + \frac{\lambda}{2}\|A_u y - q\|^2.$$

A quick computation shows

$$\begin{aligned} \nabla_y \varphi(u, y) &= \nabla f(y) + \lambda A_u^T (A_u y - q), \\ \nabla_u \varphi(u, y) &= \lambda G(u, y)^T (A_u y - q), \end{aligned} \tag{5}$$

where $G(u, y)$ is the Jacobian with respect to u of the map $u \mapsto A_u y$. Clearly the Lipschitz constant $\mathrm{Lip}(\nabla\varphi)$ scales with λ. This can be detrimental to numerical methods. For example, basic gradient descent will find a point (u_k, y_k) satisfying $\|\nabla\varphi(u_k, y_k)\|^2 < \epsilon$ after at most $\mathcal{O}\left(\frac{\mathrm{Lip}(\nabla\varphi)(\varphi(u_0,x_0)-\varphi^*)}{\epsilon}\right)$ iterations [16, Section 1.2.3].

As discussed in the introduction, minimizing φ amounts to the minimization problem $\min_u \widetilde{\varphi}(u)$ for the reduced function $\widetilde{\varphi}$ defined in (4). Note since f is convex and A_u is invertible, the function $\varphi(u, \cdot)$ admits a unique minimizer, which we denote by y_u. Appealing to the classical implicit function theorem (e.g. [20, Theorem 10.58]) we deduce that $\widetilde{\varphi}$ is differentiable with

$$\nabla\widetilde{\varphi}(u) = \nabla_u \varphi(\cdot, y_u)\Big|_u = \lambda G(u, y_u)^T (A_u y_u - q).$$

We aim to upper bound the Lipschitz constant of $\nabla\widetilde{\varphi}$ by a quantity independent of λ. We start by estimating the residual $\|A_u y_u - q\|$.

Throughout the paper, we use the following simple identity. Given an invertible map $F\colon \mathbb{R}^n \to \mathbb{R}^n$ and invertible matrix C, for any points $x \in \mathbb{R}^n$ and nonzero $\lambda \in \mathbb{R}$, we have

1) $F^{-1}(\lambda x) = (\lambda^{-1}F)^{-1}(x)$, and
2) $C \circ F^{-1} \circ C^T = \left(C^{-T} \circ F \circ C^{-1}\right)^{-1}$.

We often apply this observation to the invertible map $F(x) = \nabla f(x) + Bx$, where B is a positive definite matrix.

Lemma 1 (Residual bound)**.** *For any point u, the inequality holds:*

$$\|A_u y_u - q\| \le \frac{\|\nabla(f \circ A_u^{-1})(q)\|}{\lambda}.$$

Proof. Note that the minimizers y_u of $\varphi(u, \cdot)$ are characterized by first order optimality conditions

$$0 = \nabla f(y) + \lambda \cdot A_u^T (A_u y - q). \tag{6}$$

Applying the implicit function theorem, we deduce that y_u depends C^2-smoothly on u with $\nabla_u y_u$ given by

$$-\left(\nabla^2 f(y_u) + \lambda A_u^T A_u\right)^{-1} \nabla_u \left(\lambda A(\cdot)^T (A(\cdot)y_u - q)\right)(u).$$

On the other hand, from the equality (6) we have

$$y_u = (\nabla f + \lambda A_u^T A_u)^{-1}(\lambda A_u^T q) = \left(\frac{\nabla f}{\lambda} + A_u^T A_u\right)^{-1} A_u^T q.$$

Therefore, we deduce

$$\begin{aligned} A_u y_u - q &= A_u \left(\frac{1}{\lambda}\nabla f + A_u^T A_u\right)^{-1} A_u^T q - q \\ &= \left(\left(\frac{1}{\lambda} A_u^{-T} \circ \nabla f \circ A_u^{-1} + I\right)^{-1} - I\right) q. \end{aligned}$$

Define now the operator

$$F := \frac{1}{\lambda} A_u^{-T} \circ \nabla f \circ A_u^{-1} + I$$

and the point $z := F(q)$. Note that

$$F(x) - x = \frac{1}{\lambda}\nabla(f \circ A_u^{-1})(x).$$

Letting L be a Lipschitz constant of F^{-1}, we obtain

$$\begin{aligned}\|A_u y_u - q\| &= \|F^{-1}(q) - F^{-1}(z)\| \le L\|q - z\| \\ &= L\|q - F(q)\| = \frac{L}{\lambda}\|A_u^{-T}\nabla f(A_u^{-1}q)\|.\end{aligned}$$

Now the inverse function theorem yields for any point y the inequality

$$\begin{aligned}\|\nabla F^{-1}(y)\| &= \|\nabla F(F^{-1}(y))^{-1}\| \\ &= \left\|\left(\frac{1}{\lambda}\nabla^2(f \circ A_u^{-1})(F^{-1}(y)) + I\right)^{-1}\right\| \le 1,\end{aligned}$$

where the last inequality follows from the fact that by convexity of $f \circ A_u^{-1}$ all eigenvalues of $\nabla^2(f \circ A_u^{-1})$ are nonnegative. Thus we may set $L = 1$, completing the proof. □

For ease of reference, we record the following direct corollary.

Corollary 1. *For any point u, we have*

$$\|y_u - A_u^{-1}q\| \le \|A_u^{-1}\|\|A_u y_u - q\| \le \frac{\|A_u^{-1}\|\|\nabla(f \circ A_u^{-1})(q)\|}{\lambda}.$$

Next we will compute the Hessian of $\varphi(u, y)$, and use it to show that the norm of the Hessian of $\widetilde{\varphi}$ is bounded by a constant independent of λ. Defining

$$R(u, y, v) = \nabla_u\left[G(u, y)^T v\right] \quad \text{and} \quad K(u, v) = \nabla_u\left[A_u^T v\right],$$

we can partition the Hessian as follows:

$$\nabla^2\varphi = \begin{bmatrix}\varphi_{uu} & \varphi_{uy} \\ \varphi_{yu} & \varphi_{yy}\end{bmatrix}$$

where

$$\begin{aligned}\varphi_{uu}(u, y) &= \lambda\left(G(u, y)^T G(u, y) + R(u, y, A_u y - q)\right), \\ \varphi_{yy}(u, y) &= \nabla^2 f(y) + \lambda A_u^T A_u, \\ \varphi_{yu}(u, y) &= \lambda\left(K(u, A_u y - q) + A_u^T G(u, y)\right).\end{aligned}$$

See [22, Section 4] for more details. Moreover, it is known that the Hessian of the reduced function $\widetilde{\varphi}$ admits the expression [22, Equation 22]

$$\nabla^2\widetilde{\varphi}(u) = \varphi_{uu}(u, y_u) - \varphi_{uy}(u, y_u)\varphi_{yy}(u, y_u)^{-1}\varphi_{yu}(u, y_u), \tag{7}$$

which is simply the Schur complement of $\varphi_{yy}(u, y_u)$ in $\nabla^2\varphi(u, y_u)$. We define the operator norms

$$\begin{aligned}\left\|\nabla_u G(u, y)^T\right\| &:= \sup_{\|v\| \le 1}\left\|\nabla_u\left[G(u, y)^T v\right]\right\|, \\ \left\|\nabla_u A_u^T\right\| &:= \sup_{\|v\| \le 1}\left\|\nabla_u\left[A_u^T v\right]\right\|.\end{aligned}$$

Using this notation, we can prove the following key bounds.

Corollary 2. *For any points u and y, the inequalities hold:*

$$\begin{aligned}\|\varphi_{yy}(u, y)^{-1}\| &\le \frac{\|A_u^{-1}\|^2}{\lambda}, \\ \|R(u, y_u, A_u y_u - q)\| &\le \frac{\|\nabla(f \circ A_u^{-1})(q)\|\,\|\nabla_u G(u, y_u)\|}{\lambda}, \\ \|K(u, A_u y_u - q)\| &\le \frac{\|\nabla(f \circ A_u^{-1})(q)\|\left\|\nabla_u A_u^T\right\|}{\lambda}.\end{aligned}$$

Proof. The first bound follows by the inequality

$$\begin{aligned}\|\varphi_{yy}(u, y)^{-1}\| &= \frac{1}{\lambda}\left\|A_u^{-1}\left(\frac{1}{\lambda}A_u^{-T}\nabla^2 f(y)A_u^{-1} + I\right)^{-1}A_u^{-T}\right\| \\ &\le \frac{\|A_u^{-1}\|^2}{\lambda},\end{aligned}$$

and the remaining bounds are immediate from Lemma 1. □

Next, we need the following elementary linear algebraic fact.

Lemma 2. *For any positive semidefinite matrix B and a real $\lambda > 0$, we have $\|I - \left(I + \frac{1}{\lambda}B\right)^{-1}\|_2 \le \frac{\|B\|}{\lambda}$.*

Proof. Define the matrix $F = I - \left(I + \frac{1}{\lambda}B\right)^{-1}$ and consider an arbitrary point z. Observing the inequality $\|\left(I + \frac{1}{\lambda}B\right)^{-1}\| \le 1$ and defining the point $p := (I + \frac{1}{\lambda}B)z$, we obtain

$$\begin{aligned}\|Fz\| &= \left\|z - \left(I + \frac{1}{\lambda}B\right)^{-1}z\right\| \\ &= \left\|\left(I + \frac{1}{\lambda}B\right)^{-1}p - \left(I + \frac{1}{\lambda}B\right)^{-1}z\right\| \\ &\le \left\|\left(I + \frac{1}{\lambda}B\right)^{-1}\right\|\|p - z\| \\ &\le \left\|\frac{1}{\lambda}Bz\right\| \le \frac{\|B\|}{\lambda}\|z\|.\end{aligned}$$

Since this holds for all z, the result follows. □

Putting all the pieces together, we can now prove the main theorem of this section.

Theorem 1 (Norm of the reduced Hessian)**.** *The operator norm of $\nabla^2\widetilde{\varphi}(u)$ is bounded by a quantity $C(A(\cdot), f, q, u)$ independent of λ.*

Proof. To simplify the proof, define $G := G(u, y_u)$, $R := R(u, y_u, A_u y - q)$, $K := K(u, A_u y_u - q)$, and $\Delta = \varphi_{yy}(u, y_u)$. After rearranging (7), we can write $\nabla^2\widetilde{\varphi}$ as follows:

$$\begin{aligned}&\lambda R - \lambda^2\left(K^T\Delta^{-1}K + K^T\Delta^{-1}A_u^T G + G^T A_u\Delta^{-1}K\right) \\ &+ \lambda G^T G - \lambda^2 G^T A_u\Delta^{-1}A_u^T G.\end{aligned} \tag{8}$$

Corollary 2 implies that the operator norm of the first row of (8) is bounded above by the quantity

$$\begin{aligned}&L_u\|\nabla_u G\| + \frac{1}{\lambda}L_u^2\|\nabla_u A_u^T\|^2\|A_u^{-1}\|^2 \\ &\qquad + 2L_u\|\nabla_u A_u^T\|\|A_u^{-1}\|^2\|A_u^T G\|,\end{aligned} \tag{9}$$

where we set $L_u := \|\nabla(f \circ A_u^{-1})(q)\|$. Notice that the expression in (9) is independent of λ. We rewrite the second row of (8) using the explicit expression for Δ:

$$\begin{aligned}
&\lambda G^T G - \lambda^2 G^T A_u \Delta^{-1} A_u^T G \\
&\quad = \lambda \left(G^T G - G^T \left(\frac{1}{\lambda} A_u^{-T} \nabla^2 f(y_u) A_u^{-1} + I \right)^{-1} G \right) \\
&\quad = \lambda \left(G^T \left(I - \left(\frac{1}{\lambda} A_u^{-T} \nabla^2 f(y_u) A_u^{-1} + I \right)^{-1} \right) G \right).
\end{aligned}$$

Applying Lemma 2 with $B = A_u^{-T} \nabla^2 f(y_u) A_u^{-1}$, we have

$$\left\| \lambda G^T G - \lambda^2 G^T A_u \Delta^{-1} A_u^T G \right\| \leq \|G\|^2 \|A_u^{-T} \nabla^2 f(y_u) A_u^{-1}\|.$$

Setting

$$\begin{aligned}
C(A(\cdot), f, q, u) := L_u \|\nabla_u G\| + \frac{1}{\lambda} L_u^2 \|\nabla_u A_u^T\|^2 \|A_u^{-1}\|^2 \\
+ 2L_u \|\nabla_u A_u^T\| \|A_u^{-1}\|^2 \|A_u^T G\| \\
+ \|G\|^2 \|A_u^{-1}\|^2 \|\nabla^2 f(y_u)\|,
\end{aligned}$$

the result follows. □

A. Inexact analysis of the projection subproblem

In practice, one can rarely evaluate $\widetilde{\varphi}(u)$ exactly. It is therefore important to understand how inexact solutions of the inner problems (4) impact iteration complexity of the outer problem. The results presented in the previous section form the foundation for such an analysis. For simplicity, we assume that g is smooth, though the results can be generalized, as we comment on shortly.

In this section, we compute the overall complexity of the partial minimization technique when the outer nonconvex minimization problem (3) is solved by an inexact gradient descent algorithm. When g is nonsmooth, a completely analogous analysis applies to the prox-gradient method. We only focus here on gradient descent, as opposed to more sophisticated methods, since the analysis is straightforward. We expect quasi-Newton methods and limited memory variants to exhibit exactly the same behavior (e.g. Figure 1). We do not perform a similar analysis here for inexact quasi-Newton methods, as the global efficiency estimates even for exact quasi-Newton methods for nonconvex problems are poorly understood.

Define the function $H(u) := g(u) + \widetilde{\varphi}(u)$. Let $\beta > 0$ be the Lipschitz constant of the gradient $\nabla H = \nabla g + \nabla \widetilde{\varphi}$. Fix a constant $c > 0$, and suppose that in each iteration k, we compute a vector v_k with $\|v_k - \nabla H(u_k)\| \leq \frac{c}{k}$. Consider then the inexact gradient descent method $u_{k+1} = u_k - \frac{1}{\beta} v_k$. Then we deduce

$$\begin{aligned}
H(u_{k+1}) - H(u_k) &\leq -\langle \nabla H(u_k), \beta^{-1} v_k \rangle + \frac{\beta}{2} \|\beta^{-1} v_k\|^2 \\
&= \frac{1}{2\beta} \left(\|v_k - \nabla H(u_k)\|^2 - \|\nabla H(u_k)\|^2 \right).
\end{aligned} \tag{10}$$

Hence we obtain the convergence guarantee:

$$\begin{aligned}
\min_{i=1,\ldots,k} \|\nabla H(u_i)\|^2 &\leq \frac{1}{k} \sum_{i=1}^k \|\nabla H(u_i)\|^2 \\
&\leq \frac{2\beta \left(H(u_1) - H^* \right)}{k} + \frac{1}{k} \sum_{i=1}^k \|v_k - \nabla H(u_k)\|^2 \\
&\leq \frac{2\beta \left(H(u_1) - H^* \right)}{k} + \frac{c^2 \pi^2}{6k} \leq \frac{\beta^2 \left\| u_1 - u^* \right\|^2 + c^2 \pi^2 / 6}{k}.
\end{aligned}$$

where (10) is used to go from line 1 to line 2. Now, if we compute ∇g exactly, the question is how many inner iterations are needed to guarantee $\|v_k - \nabla H(u_k)\| \leq \frac{c}{k}$. For fixed $u = u_k$, the inner objective is

$$\varphi(u_k, y) = f(y) + \frac{\lambda}{2} \|A(u_k) y - q\|^2.$$

The condition number (ratio of Lipschitz constant of the gradient over the strong convexity constant) of $\varphi(u_k, y)$ in y is

$$\kappa_k := \frac{1}{\lambda} \mathrm{Lip}(f) \|A(u_k)^{-1}\|^2 + \|A(u_k)\|^2 \|A(u_k)^{-1}\|^2.$$

Notice that κ_k converges to the squared condition number of $A(u_k)$ as $\lambda \uparrow \infty$. Gradient descent on the function $\varphi(u_k, \cdot)$ guarantees $\|y_i - y^*\|^2 \leq \epsilon$ after $\kappa_k \log \left(\frac{\|y_0 - y^*\|^2}{\epsilon} \right)$ iterations. Then we have

$$\|\nabla_u \varphi(u_k, y_i) - \nabla_u \varphi(u_k, y^*)\| \leq \lambda \|\nabla_u G(u, y^*)\| \|y_i - y^*\|.$$

Since we want the left hand side to be bounded by $\frac{c}{k}$, we simply need to ensure

$$\|y_i - y^*\|^2 \leq \frac{c^2}{k^2 \lambda^2 \|\nabla_u G(u_k, y^*)\|^2}.$$

Therefore the total number of inner iterations is no larger than

$$\kappa_k \log \left(\frac{\|y_0 - y^*\|^2 \|\nabla_u G(u_k, y^*)\|^2}{c^2} k^2 \lambda^2 \right),$$

which grows very slowly with k and with λ. In particular, the number of iterations to solve the inner problem scales as $\log(k\lambda)$ to achieve a global $\frac{1}{k}$ rate in $\|\nabla H\|^2$. If instead we use a fast-gradient method [16, Section 2.2] for minimizing $\varphi(u_k, \cdot)$, we can replace κ_k with the much better quantity $\sqrt{\kappa_k}$ throughout.

III. Numerical Illustrations

In this section, we present two representative examples of PDE constrained optimization (boundary control and optimal transport) and a problem of robust dynamic inference. In each case, we show that practical experience supports theoretical results from the previous section. In particular, in each numerical experiment, we study the convergence behavior of the proposed method as λ increases.

A. Boundary control

In boundary control, the goal is to steer a system towards a desired state by controlling its boundary conditions. Perhaps the simplest example of such a problem is the following. Given a source $q(x)$, defined on a domain Ω, we seek boundary conditions u such that the solution to the Poisson problem

$$\begin{aligned} \Delta y &= q \quad \text{for } x \in \Omega \\ y|_{\partial\Omega} &= u \end{aligned}$$

is close to a desired state y_d. Discretizing the PDE yields the system

$$Ay + Bu = q$$

where A is a discretization of the Laplace operator on the interior of the domain and B couples the interior gridpoints to the boundary. The corresponding PDE-constrained optimization problem is given by

$$\min_{u,y} \ \tfrac{1}{2}\|y - y_d\|_2^2 \qquad \text{subject to} \qquad Ay + Bu = q,$$

whereas the penalty formulation reads

$$\min_{u,y} \ \tfrac{1}{2}\|y - y_d\|_2^2 + \tfrac{\lambda}{2}\|Ay + Bu - q\|_2^2.$$

Since both terms are quadratic in y, we can quickly solve for y explicitly.

1) Numerical experiments: In this example, we consider an L-shaped domain with a source q shaped like a Gaussian bell, as depicted in figure 2. Our goal is to get a constant distribution $y_d = 1$ in the entire domain. The solution for $u = 1$ is shown in figure 3 (a). To solve the optimization problem we use a steepest-descent method with a fixed step-size, determined from the Lipschitz constant of the gradient. The result of the constrained formulation is shown in figure 3 (b). We see that by adapting the boundary conditions we get a more even distribution. The convergence behavior for various values of λ is shown in figure 4 (a). We see that as $\lambda \uparrow \infty$, the behaviour tends towards that of the constrained formulation, as expected. The Lipschitz constant of the gradient (evaluated at the initial point u), as a function of λ is shown in figure 4 (b); the curve levels off as the theory predicts.

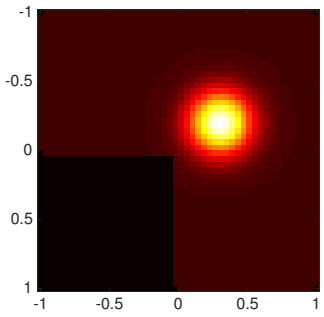
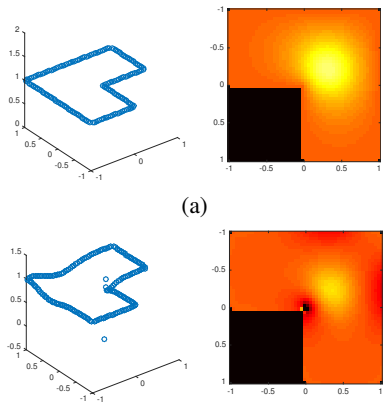


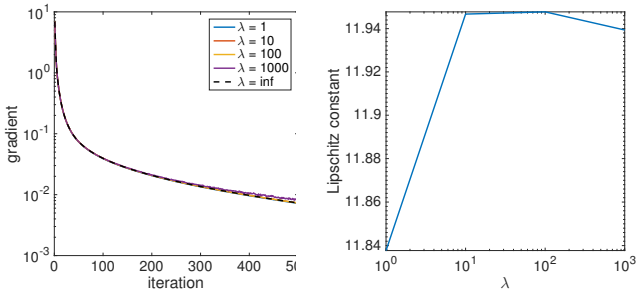
Fig. 2: L-shaped domain with the source function q.



(a)

(b)

Fig. 3: Boundary values, u, and solution in the interior for the initial and optimized boundary values are depicted in (a) and (b) respectively.



(a) (b)

Fig. 4: The convergence plots for various values of λ are depicted in (a), while (b) shows the dependence of the (numerically computed) Lipschitz constant on λ.

B. Optimal transport

The second class of PDE-constrained problems we consider comes from optimal transport, where the goal is to determine a mapping, or flow, that optimally transforms one mass density function into another. Say we have two density functions $y_0(x)$ and $y_T(x)$, with $x \in \Omega$, we can formulate the problem as finding a flowfield, $u(t,x) = \begin{pmatrix} u_1(t,x) \\ u_2(t,x) \end{pmatrix}$, such that $y_T(x) =$

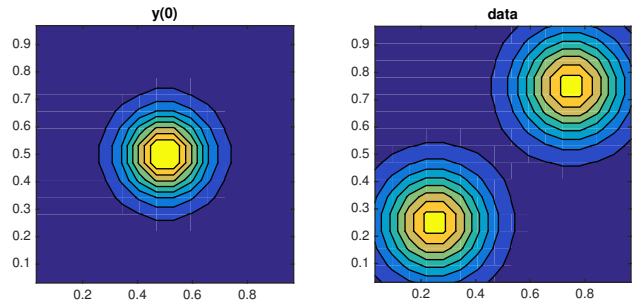


Fig. 5: The initial (left) and desired (right) mass density are shown.

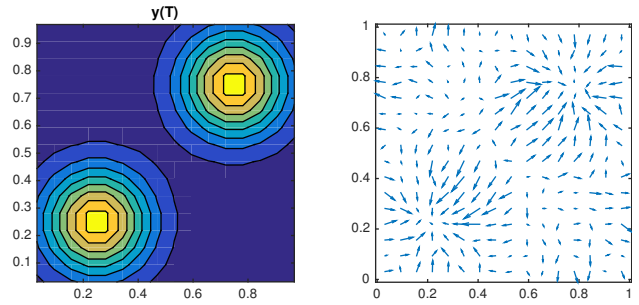
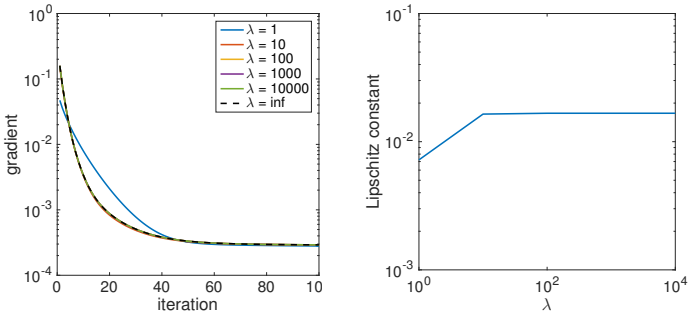


Fig. 6: The mass density obtained after optimization (left) and the corresponding time-averaged flow field (right) are shown.



(a) (b)

Fig. 7: Convergence plots for various values of λ are shown in (a), while (b) shows the (numerically computed) Lipschitz constant as a function of λ.

$y(T,x)$ and $y_0(x) = y(0,x)$, where $y(t,x)$ solves

$$y_t + \nabla \cdot (yu) = 0.$$

Discretizing using an implicit Lax-Friedrichs scheme [12], the PDE reads

$$A(u)y = q,$$

where q contains the initial condition and we have

$$A(u) = \begin{pmatrix} I + \Delta t B(u^1) & & & \\ -M & I + \Delta t B(u^2) & & \\ & -M & & \\ & & & -M \quad I + \Delta t B(u^N) \end{pmatrix},$$

with M a four-point averaging matrix and B containing the discretization of the derivative terms. Adding regularization to promote smoothness of u and y in time [12], we obtain the problem

$$\begin{aligned} \min_{u,y} \ & \tfrac{1}{2}\|Py - y_T\|_2^2 + \tfrac{\alpha^2}{2} y^T L \mathrm{diag}(u) u \\ \text{subject to} \quad & A(u)y = q. \end{aligned} \tag{11}$$

Here, P restricts the solution y to $t = T$, α is a regularization parameter and L is a block matrix with I on the main and upper diagonal. The penalized formulation is

$$\min_{u,y} \ \tfrac{1}{2}\|Py - y_T\|_2^2 + \tfrac{\alpha^2}{2} y^T L \mathrm{diag}(u) u + \tfrac{1}{2}\lambda \|A(u)y - q\|^2. \tag{12}$$

Again the partial minimization in y amount to minimizing a quadratic function.

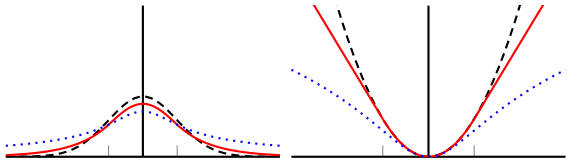


Fig. 8: Left: Densities, Gaussian (black dash), Huber (red solid), and Student's t (blue dot). Right: Negative Log Likelihoods.

1) Numerical experiments: For the numerical example we consider the domain $\Omega = [0,1]^2$, discretized with $\Delta x = 1/16$ and $T = 1/32$ with a stepsize of $\Delta t = 1/8$. The initial and desired state are depicted in figure 5. The resulting state obtained at time T and the corresponding time-averaged flowfield are depicted in figure 6. The initial flow u_0 was generated by i.i.d. samples from a standard Gaussian random variable. To minimize (12), we used a steepest-descent method with constant step size, using the largest eigenvalue of the Gauss-Newton Hessian at the initial u as an estimate of the Lipschitz constant. The convergence behavior for various values of λ as well as the corresponding estimates of the Lipschitz constant at the final solution are shown in figure 7.

C. Robust dynamic inference with the penalty method

In many settings, data is naturally very noisy, and a lot of effort must be spent in pre-processing and cleaning before applying standard inversion techniques.

To narrow the scope, consider dynamic inference, where we wish to infer both hidden states and unknown parameters driven by an underlying ODE. Recent efforts have focused on developing inference formulations that are robust to outliers in the data [3], [4], [9], using convex penalties such as ℓ_1, Huber [14] and non-convex penalties such as the Student's t log likelihood in place of the least squares penalty. The goal is to develop formulations and estimators that achieve adequate performance when faced with outliers; these may arise either as gross measurement errors, or real-world events that are not modeled by the dynamics.

Figure 8 shows the probability density functions and penalties corresponding to Gaussian, Huber, and Student's t densities. Quadratic tail growth corresponds to extreme decay of the Gaussian density for large inputs, and linear growth of the influence of any measurement on the fit. In contrast, Huber and Student's t have linear and sublinear tail growth, respectively, which ensures every observation has bounded influence.

We focus on the Huber function [14], since it is both C^1-smooth and convex. In particular, the function $f(y)$ in (1) and (4) is chosen to be a composition of the Huber with an observation model. Note that Huber is not C^2, so this case is not immediately captured by the theory we propose. However, Huber can be closely approximated by a C^2 function [8], and then the theory fully applies. For our numerical examples, we apply the algorithm developed in this paper directly to the Huber formulation.

We illustrate robust modeling using a simple representative example. Consider a 2-dimensional oscillator, governed by the

following equations:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}' = \begin{bmatrix} -2u_1u_2 & -u_1^2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} \sin(\omega t) \\ 0 \end{bmatrix} \quad (13)$$

where we can interpret u_1 as the frequency, and u_2 is the damping. Discretizing in time, we have

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}^{k+1} = \begin{bmatrix} 1 - 2\Delta t u_1 u_2 & -\Delta t u_1^2 \\ \Delta t & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}^{k} + \Delta t \begin{bmatrix} \sin(\omega t_k) \\ 0 \end{bmatrix}.$$

We now consider direct observations of the second component,

$$z_k = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}^{k} + w_k, \quad w_k \sim N(0, R_k).$$

We can formulate the joint inference problem on states and measurements as follows:

$$\min_{u,y} \ \phi(u,y) := \frac{1}{2}\rho\left(R^{-1/2}(Hy - z)\right) \ \text{ s.t. } \ Gy = v, \quad (14)$$

with $v_k = \sin(\omega t_k)$, v_1 the initial condition for y, and ρ is either the least squares or Huber penalty, and we use the following definitions:

$$\begin{aligned} y &= \text{vec}(\{y_k\}) \\ z &= \text{vec}(\{z_1, \ldots, z_N\}) \\ R = \text{diag}(\{R_k\}) & \\ H = \text{diag}(\{H_k\}) & \quad G = \begin{bmatrix} \mathrm{I} & 0 & & \\ -G_2 & \mathrm{I} & \ddots & \\ & \ddots & \ddots & 0 \\ & & -G_N & \mathrm{I} \end{bmatrix} \end{aligned}$$

Note in particular that there are only two unknown parameters, i.e. $u \in \mathbb{R}^2$, while the state y lies in $\mathbb{R}^{2N}$, with N the number of modeled time points.

The reduced optimization problem for u is given by

$$\min_u f(u) := \frac{1}{2}\rho(R^{-1/2}(HG(u)^{-1}v - z).$$

To compute the derivative of the ODE-constrained problem, we can use the adjoint state method. Defining the Lagrangian

$$\mathcal{L}(y,x,u) = \frac{1}{2}\rho\left(R^{-1/2}(Hy - z)\right) + \langle x, Gy - v\rangle,$$

we write down the optimality conditions $\nabla\mathcal{L} = 0$ and obtain

$$\left\{ \begin{array}{l} \quad \overline{y} = G^{-1}v \\ \quad \overline{x} = -G^{-T}(H^T R^{-1/2} \nabla\rho(R^{-1/2}HG^{-1}v - z)) \\ \nabla_u f = \left\langle \overline{x}, \dfrac{\partial(G(u)\overline{y})}{\partial u} \right\rangle \end{array} \right\}.$$

The inexact (penalized) problem is given by

$$\min_{u,y} \ \varphi(u,y) = \frac{1}{2}\rho\left(R^{-1/2}(Hy - z)\right) + \frac{\lambda}{2}\|Gy - v\|^2. \quad (15)$$

We then immediately find

$$\begin{aligned} \overline{y} &= \arg\min_y \frac{1}{2}\rho\left(R^{-1/2}(Hy - z)\right) + \frac{\lambda}{2}\|Gy - v\|^2 \\ \nabla\widetilde{\phi}(u) &= \lambda\frac{\partial(G(u)\overline{y})}{\partial u}(G\overline{y} - v). \end{aligned}$$

When ρ is the least squares penalty, $\overline{y}$ is available in closed form. However, when ρ is the Huber, $\overline{y}$ requires an iterative algorithm.

TABLE I: Results for Kalman experiment. Penalty method for both least squares and huber achieves the same results for moderate values of λ as does the projected formulation. While Huber results converge to nearly the true parameters u, least squares results converge to an incorrect parameter estimate.

λ	ρ	Iter	Opt	u
10^3	ℓ_2	9	3.1×10^{-7}	$(.45, .98)$
10^5	ℓ_2	18	3.2×10^{-7}	$(.15, 4.3)$
10^7	ℓ_2	26	5×10^{-5}	$(.07, 11.1)$
10^9	ℓ_2	31	4×10^{-6}	$(.07, 11.8)$
∞	ℓ_2	29	3.3×10^{-7}	$(.07, 11.8)$
10^3	h	9	2.4×10^{-7}	$(1.92.14)$
10^5	h	12	5×10^{-6}	$(1.98, .11)$
10^7	h	10	5×10^{-6}	$(1.99, .11)$
10^9	h	13	1×10^{-5}	$(1.99, .11)$
∞	h	17	2×10^{-6}	$(1.99, .11)$

Rather than solving for $\overline{y}$ using a first-order method, we use IPsolve, an interior point method well suited for Huber [5]. Even though each iteration requires inversions of systems of size $\mathcal{O}(N)$, these systems are very sparse, and the complexity of each iteration to compute $\overline{y}$ is $\mathcal{O}(N)$ for any piecewise linear quadratic function [5]. Once again, we see that the computational cost does not scale with λ.

1) Numerical Experiments: We simulate a data contamination scenario by solving the ODE (13) for the particular parameter value $u = (2, 0.1)$. The second component of the resulting state y is observed, and the observations are contaminated. In particular, in addition to Gaussian noise with standard deviation $\sigma = 0.1$, for 10% of the measurements uniformly distributed errors in $[0, 2]$ are added. The state $y \in \mathbb{R}^{2(4000)}$ is finely sampled over 40 periods.

For the least squares and the Huber penalty with $\kappa = 0.1$, we solved both the ODE constrained problem (14) and the penalized version (15) for $\lambda \in \{10^3, 10^5, 10^7, 10^9\}$. The results are presented in Table I. The Huber formulation behaves analogously to the formulation using least squares; in particular the outer (projected) function in u is no more difficult to minimize. And, as expected, the robust Huber penalty finds the correct values for the parameters.

A state estimate generated from u-estimates corresponding to large λ is shown in Figure 9. The huberized approach is able to ignore the outliers, and recover both better estimates of the underlying dynamics parameters u, and the true observed and hidden components of the state y.

IV. Conclusions

In this paper, we showed that, contrary to conventional wisdom, the quadratic penalty technique can be used effectively for control and PDE constrained optimization problems, if done correctly. In particular, when combined with the partial minimization technique, we showed that the penalized projected scheme

$$\min_u \ g(u) + \min_y \left\{ f(y) + \frac{\lambda}{2}\|A(u)y - b\|^2 \right\}$$

has the following advantages:

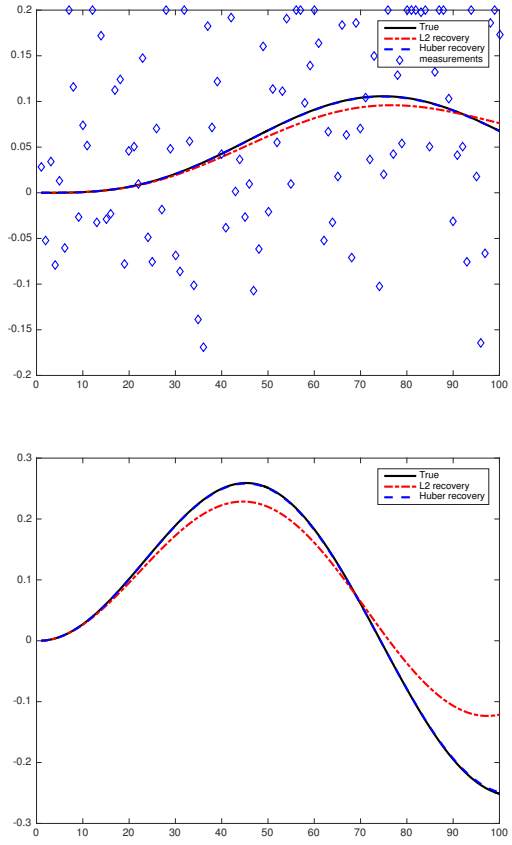


Fig. 9: Top panel: first 100 samples of state y_2 (solid black), least squares recovery (red dash-dot) and huber recovery (blue dash). Noisy measurements z appear as blue diamonds on the plot, with outliers shown at the top of the panel. Bottom panel: first 100 samples of state y_1 (solid black), least squares recovery (red dash-dot) and Huber recovery (blue dash).

1) The Lipschitz constant of the gradient of the outer function in u is bounded as $\lambda \uparrow \infty$, and hence we can effectively analyze the global convergence of first-order methods.
2) Convergence behavior of the data-regularized convex inner problem is controlled by parametric matrix $A(\cdot)$, a fundamental quantity that does not depend on λ.
3) The inner problem can be solved inexactly, and in this case, the number of inner iterations (of a first-order algorithm) needs to grow only logarithmically with λ and the outer iterations counter, to preserve the natural rate of gradient descent.

As an immediate application, we extended the penalty method in [22] to convex robust formulations, using the Huber penalty composed with a linear model as the function $f(y)$. Numerical results illustrated the overall approach, including convergence behavior of the penalized projected scheme, as well as modeling advantages of robust penalized formulations.

Acknowledgment. Research of A. Aravkin was partially supported by the Washington Research Foundation Data Science Professorship. Research of D. Drusvyatskiy was partially supported by the AFOSR YIP award FA9550-15-1-0237. The research of T. van Leeuwen was in part financially supported by the Netherlands Organisation of Scientific Research (NWO) as part of research programme 613.009.032.

REFERENCES

[1] L. Ambrosio and N. Gigli. A User's Guide to Optimal Transport. In *Modelling and Optimisation of Flows on Networks*, pages 1–155. Springer, 2013.
[2] B. D. Anderson and J. B. Moore. Optimal filtering. 1979, 1979.
[3] A. Aravkin, B. Bell, J. Burke, and G. Pillonetto. An ℓ_1-Laplace robust Kalman smoother. *IEEE Transactions on Automatic Control*, 2011.
[4] A. Aravkin, J. Burke, and G. Pillonetto. Robust and trend-following Student's t Kalman smoothers. *SIAM Journal on Control and Optimization*, 52(5):2891–2916, 2014.
[5] A. Aravkin, J. V. Burke, and G. Pillonetto. Sparse/robust estimation and Kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory. *Journal of Machine Learning Research*, 14:2689–2728, 2013.
[6] A. Aravkin, M. Friedlander, F. Herrmann, and T. van Leeuwen. Robust inversion, dimensionality reduction, and randomized sampling. *Mathematical Programming*, 134(1):101–125, 2012.
[7] A. Y. Aravkin and T. van Leeuwen. Estimating nuisance parameters in inverse problems. *Inverse Problems*, 28(11):115016, 2012.
[8] K. P. Bube and R. T. Langan. Hybrid $\ell 1/\ell 2$ minimization with applications to tomography. *Geophysics*, 62(4):1183–1195, 1997.
[9] S. Farahmand, G. B. Giannakis, and D. Angelosante. Doubly Robust Smoothing of Dynamical Processes via Outlier Sparsity Constraints. *IEEE Transactions on Signal Processing*, 59:4529–4543, 2011.
[10] G. Golub and V. Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Inverse Problems*, 19(2):R1–R26, Apr. 2003.
[11] M. Gugat. Control problems with the wave equation. *SIAM Journal on Control and Optimization*, 48(5):3026–3051, 2009.
[12] E. Haber and L. Hanson. Model Problems in PDE-Constrained Optimization. Technical report, 2007.
[13] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent. Optimal Mass Transport for Registration and Warping. *International Journal of Computer Vision*, 60(3):225–240, dec 2004.
[14] P. J. Huber. *Robust Statistics*. John Wiley and Sons, 2004.
[15] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the AMSE - Journal of Basic Engineering*, 82(D):35–45, 1960.
[16] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.
[17] J. Nocedal and S. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.
[18] C. C. Paige and M. A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS)*, 8(1):43–71, 1982.
[19] P. Philip. Optimal Control and Partial Differential Equations. Technical report, LMU Berlin, 2009.
[20] R. Rockafellar and R. Wets. *Variational Analysis*, volume 317. Springer, 1998.
[21] A. Tarantola. *Inverse Problem Theory*. SIAM, 2005.
[22] T. van Leeuwen and F. J. Herrmann. A penalty method for pde-constrained optimization in inverse problems. *Inverse Problems*, 32(1):015007, 2015.
[23] J. Virieux and S. Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, 2009.