Variable projection without smoothness

A.Y. Aravkin[1], D. Drusvyatskiy[1], and T. van Leeuwen[2]

[1]University of Washington, Seattle, WA, USA
[2]University of Utrecht, Utrecht, Netherlands

January 19, 2016

Abstract

Variable projection is a powerful technique in optimization. Over the last 30 years, it has been applied broadly, with empirical and theoretical results demonstrating both greater efficacy and greater stability than competing approaches. In this paper, we illustrate the technique on a large class of structured nonsmooth optimization problems, with numerical examples in sparse deconvolution and machine learning applications.

1 Introduction

This paper revolves around the *Variable Projection* technique in nonsmooth optimization. One of the prominent early references on the topic [11] concentrates on separable least-squares problems, having numerous applications in chemistry, mechanical systems, neural networks, and telecommunications; see the surveys of [12] and [18], and references therein. Setting the stage, consider a *separable least-squares problem*

$$\min_{x,\theta} f(x,\theta) := \|y - \Phi(x)\theta\|_2^2 \,, \tag{1}$$

where the matrix-valued map $\Phi(x)$ is smooth in the parameter x. One may formally eliminate the variable θ by rewriting the problem:

$$\min_x \ \widetilde{f}(x) \qquad \text{where} \qquad \widetilde{f}(x) := \min_\theta \ f(x,\theta). \tag{2}$$

Though $\widetilde{f}(x)$ is implicitly defined, it can be explicitly evaluated, since $f(x,\cdot)$ is a convex quadratic in θ. The variable projection technique for this class of problems, going back to [11], then aims to solve the original problem (1) by instead running a nonlinear optimization solver on $\min_x \widetilde{f}(x)$. The authors of [27] showed that when the Gauss-Newton method for (1) converges superlinearly, so do Gauss-Newton variants for (2); moreover, empirical evidence suggests that the latter schemes outperform the former.

The underlying principle is much broader than the class of nonlinear least squares problems. For example, the authors of [4, 5] consider the class of problems

$$\min_{x,\theta} \ f(x,\theta) \,, \tag{3}$$

where $f : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ is a C^2-smooth function, the vector $x \in \mathbb{R}^n$ models the primary set of parameters, while $\theta \in \mathbb{R}^k$ is a secondary set of nuisance parameters (such as variance, tuning, or regularization). Applying the variable projection technique requires approximate evaluation of the function $\widetilde{f}(x) = \min_\theta f(x,\theta)$ along with its first (and potentially second) derivatives; such formulas are readily available under mild conditions [5].

There is significant practical interest in *nonsmooth* functions f, generalizing (3). Nonsmooth regularization in both x and θ occurs often in high dimensional and signal processing settings, as well as in PDE constrained optimization, and machine learning problems. In

many cases, problem-specific projection-based algorithms have been developed. Our goal is to provide a unifying framework for a wide range of problems, illustrating its current uses in a range of fields, and to explore new applications.

The outline of the paper is as follows. In Section 2, we present important applications of nonsmooth variable projection algorithms. Throughout, we emphasize the subtleties involved, and in several cases propose novel techniques. In Section 3, we justify some technical claims made in Section 2, using basic techniques of variational analysis. Section 4 presents promising numerical results with applications to exponential fitting, sparse deconvolution, and robust learning formulations. Conclusions complete the paper.

2 Nonsmooth Variable Projection in Applications

As alluded to in the introduction, the variable projection technique has been widely used in different application domains. In this section, we describe a number of notable examples: PDE-constrained optimization, exponential data fitting, robust inference, and multiple kernel learning. We emphasize that despite the variety of these applications, the underlying algorithmic techniques all fall within the same paradigm. However, there are subtle modifications that may be required to make the scheme work, which we emphasize when appropriate. As a warm-up for the general technique considered, we begin with ODE or PDE constrained problems.

2.1 PDE-constrained optimization

Optimization problems with ODE and PDE constraints occur often when modeling physical systems; see e.g. optimal control [10,22] and inverse problems in geophysics [6,13] astrophysics [14] and medical imaging [1]. Setting notation, consider the problem

$$\min_{x,u} \; g(x,u) \quad \text{s.t.} \quad c(x,u) = 0, \tag{4}$$

where u is the *state* variable (a vector field), the system $c(x,u) = 0$ is an ODE or PDE with parameters of interest x, and g is an objective that we aim to minimize (e.g. data misfit). Here we assume that g is lower bounded and convex. Upon discretization, u becomes a high-dimensional vector. For the current exposition, we will assume that the PDE is linear. Abusing notation slightly, we assume that the system (4) is already discretized, and is in the form

$$\min_{x,u} \; g(x,u) \quad \text{s.t.} \quad H(x)u = q, \tag{5}$$

where $H(x)$ is a square invertible matrix depending C^1-smoothly on x, the function g is C^1-smooth and the margin function $g(x,\cdot)$ is convex for each x.

In large-scale applications, it is impractical to store and update the entire vector u. Instead, *adjoint state methods* compute u on the fly and optimize in x alone. Observe that the problem is equivalent to minimizing the smooth function $\widetilde{f}(x) := g(x, H(x)^{-1}q)$ and it is well-known [13,23] that $\nabla\widetilde{f}(x)$ is given by

$$\nabla\widetilde{f}(x) = \nabla_x g(x,\bar{u}) + \nabla\langle\bar{v}, H(\cdot)\bar{u}\rangle(x),$$

where $(\bar{u},\bar{v})$ satisfy the equations

$$\left\{ \begin{array}{rl} H(x)^*\bar{v} + \nabla_u g(x,\bar{u}) & = 0 \\ H(x)\bar{u} & = q \end{array} \right\}. \tag{6}$$

A variational viewpoint elucidates the formula. Indeed, basic convex duality (e.g. [7, Corollary 3.3.11]) implies that $\widetilde{f}(x)$ can equivalently be written as

$$\widetilde{f}(x) = \sup_v \min_u \; g(x,u) + \langle v, H(x)u - q\rangle. \tag{7}$$

Now observe that equation (6) simply says that $(\bar{u},\bar{v})$ is a saddle point of (7) while the description of $\nabla\widetilde{f}(x)$ is the partial gradient with respect to x of the inner function in (7) evaluated at $(x,\bar{u},\bar{v})$.

Penalty methods for PDE constrained optimization In some applications, it may be desirable to relax the constraint $H(x)u = q$ and formulate a penalized problem introduced in [35]

$$\min_{x,u} \; g(u,x) + \lambda \cdot p\Big(H(x)u - q\Big),$$

where λ is a penalty parameter and p is strongly convex, smooth penalty function (e.g. $p = \frac{1}{2}\|\cdot\|_2^2$). The variable projection technique is immediate here. An easy application of [25][Theorem 10.58] shows that the function

$$\widetilde{g}(x) := \min_{u} \; g(u,x) + \lambda \cdot p\Big(H(x)u - q\Big) \tag{8}$$

is differentiable and its derivative is simply the partial derivative of the inner function in x evaluated at $(x, \bar{u})$, where $\bar{u}$ achieves the minimum. For an extended explanation, see Section 3. Notice that in case of the least squares objective $g(u,x) = \|Ru - d\|^2$, only one equation solve is needed to compute $\bar{u}$ and hence the derivate $\nabla\widetilde{g}(x)$. This is in contrast to the adjoint method where another equation solve is needed to compute the adjoint variable $\bar{v}$. For promising numerical results, see [35]. Hence the variable projection technique, even in this rudimentary setting, can yield surprising computational benefits.

2.2 Exponential data-fitting

In this section, we present the general class of *exponential data-fitting* problems – one of the prime applications of variable projection [21]. The general formulation of these problems starts with a signal model of the form

$$y_i = \sum_{j=1}^{n} a_j \exp(-\varphi_{ij}(\theta)),$$

where $a \in \mathbb{R}^n$ are unknown weights, $y \in \mathbb{R}^m$ are the measurements, and φ_{ij} are given functions that depend on an unknown parameter $\theta \in \mathbb{R}^k$. Some examples of this class are given in table 1.

problem	known parameters	unknown parameters	φ_{ij}
pharmaco-kinetic	time t	decay rate θ	$\theta_i t_j$
signal classification	distances x_j	direction θ	$\imath\theta_i x_j$
radial basis functions	locations x_j	center r and scale s	$s_i^2\|x_j - r_i\|_2^2$

Table 1: Some examples of exponential data-fitting in applications.

Introducing the matrix $\Phi(\theta)$ with entries $\Phi(\theta)_{ij} := \exp(-\varphi_{ij}(\theta))$, we express the noisy measurements as

$$y = \Phi(\theta)a + \epsilon.$$

A natural approach is then to formulate a nonlinear least squares problem,

$$\min_{a,\theta} \; \|\Phi(\theta)a - y\|_2^2,$$

which is readily solved using the classic variable projection algorithm. In practice, however, we may not know the number of terms, n, to include. We can of course over-parametrize the problem, choosing a rather large value for n, but this may lead to unreliable estimates. By including a sparsity prior on a we force a sparser solution. We arrive at the formulation:

$$\min_{a,\theta} \; \|\Phi(\theta)a - y\|_2^2 + \lambda\|a\|_1.$$

Note also that we can replace the ℓ_2 data misfit term to account for the statistics of the noise. Additional constraints on a may be added as well [19, 33]. The variable projection approach allows us to efficiently solve this problem by projecting out a using any suitable method to solve LASSO problems (we use IPsolve [3]) and using a non-linear optimization

method to find the optimal θ. Indeed, we will see in Section 3 that provided $\Phi(\theta)$ has full column rank for all θ, the function

$$\widetilde{f}(\theta) = \min_a f(\theta, a) := \|\Phi(\theta)a - y\|_2^2 + \lambda\|a\|_1, \tag{9}$$

is differentiable with gradient $\nabla\widetilde{f}(\theta) = \nabla_\theta f(\theta, \bar{a})$ where $\bar{a}$ achieves the minimum. If $\Phi(\theta)$ does not have full column rank, we can add a quadratic regularizer in a and obtain an analogous result. We note that numerical optimization in both variables (θ, a), without variable projection, performs poorly in comparison to the reduced approach; see section 4 and figure 2.

2.3 Trimmed Robust Formulations in Machine Learning

Data contamination is a significant challenge for a range of inference problems. An alternative to robust penalties is the *trimmed* approach [15, 28], where one solves an inference problem by minimizing over k least residuals. Suppose that our inference problem can be written as

$$\min_x \sum_i g_i(x) + r(x),$$

where g_i represent the error or the negative log-likelihood of the i'th component, while $r(x)$ is any regularizer. The *trimmed* approach aims to only use k "best" components, treating the rest as "outliers" to be excluded. However, correct classification of these residuals can only be done once x is known. A natural approach is to use the k smallest residuals at each iteration of a numerical scheme. Recently, such a scheme was used to develop robust graphical models [37], using a sparsifying regularizer $r(x) = \lambda\|x\|_1$.

Consider the reformulation

$$\min_x \left(\min_{w \in \Delta_k \cap \mathbb{B}_\infty} \sum_i w_i g_i(x) \right) + r(x), \tag{10}$$

where $\Delta_k = \{w \geq 0 : 1^T w = k\}$ is a scaled simplex, and $\mathbb{B}_\infty = \{w : -1 \leq w_i \leq 1\}$ is the infinity norm unit ball. Variable projection naively seems directly applicable, since the function $x \mapsto \min_{w \in \Delta_k \cap \mathbb{B}_\infty} \sum_i w_i g_i(x)$ simply selects the k smallest $g_i(x)$'s. However, it is easy to see that this function is nonsmooth in the simplest of cases; see Fig. 1.

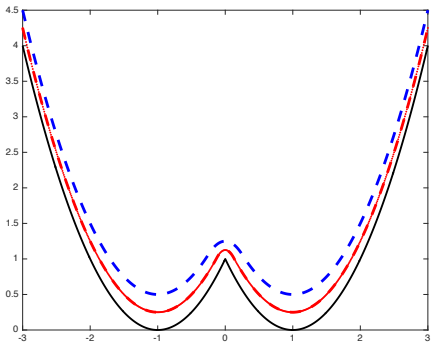


Figure 1: $g_1(x) = (x-1)^2$ and $g_2(x) = (x+1)^2$. The resulting function $\min\{g_1, g_2\}$ is nonsmooth at the origin (black solid curve). The effect of the smoothing formulation (11) shown in dashed blue for $\beta = 1$, and dash-dot red for $\beta = 0.5$.

Instead, we propose a regularized formulation:

$$\min_x \left(\min_{w \in \Delta_k} \sum_i w_i g_i(x) + \frac{\beta}{2}\|w\|^2 \right) + r(x). \tag{11}$$

We then seek to solve the problem $\min_x \widetilde{g}(x) + r(x)$, where

$$\begin{aligned}\widetilde{g}_\beta(x) &:= \min_{w \in \Delta_k \cap \mathbb{B}_\infty} \sum_i w_i g_i(x) + \frac{\beta}{2}\|w\|^2 \\ &= \frac{\beta}{2}\mathrm{dist}^2_{\Delta_k \cap \mathbb{B}_\infty}\left(-\beta^{-1} g(x)\right) - (2\beta)^{-1}\|g(x)\|^2\end{aligned}$$

Here we use the notation $g(x) := (g_1(x), \ldots, g_m(x))$. A standard computation shows that $\widetilde{g}_\beta(x)$ is indeed differentiable with gradient $\nabla \widetilde{g}_\beta(x) = \nabla g(x)\bar{x}$, where $\bar{x}$ is the nearest point to $-\beta^{-1}g(x)$ in the set of interest $\Delta_k \cap \mathbb{B}_\infty$. Projection onto this set has been recently considered and implemented [36]. Hence standard methods are directly applicable to solve (11), including prox-gradient for non-smooth $r(x)$, provided ∇g is Lipschitz continuous. In the numerical section of this paper, we illustrate the approach on a logistic regression model with data contamination. In section 3, we observe that the regularization technique just described fits into a broader paradigm.

This example illustrates the importance of a theoretical framework underlying algorithmic development. While (10) is not differentiable, a small modification (11) is, and shows good performance in applications.

2.4 Multiple Kernel Learning

Kernel methods are a powerful technique in classification and prediction [32, 34]. In such problems we are given a set of m samples $x_i \in \mathbb{R}^n$ and corresponding labels $y_i \in \mathbb{R}$ and the goal is to classify new samples. To do so, we search for a function $g(x) = \langle v, \Phi(x)\rangle \colon \mathbb{R}^n \to \mathbb{R}$, with Φ a given map from $\mathbb{R}^n$ to a specified (possibly infinite-dimensional) function class $\mathcal{H}$ (Reproducing Kernel Hilbert Space [29, 31]), such that $\langle v, \Phi(x_i)\rangle \approx y_i$. We can then use v to classify new samples. This problem can be formalized as follows

$$\begin{aligned}\min_{v \in \mathcal{H},\, b \in \mathbb{R},\, \xi \in \mathbb{R}^m_+} \quad & \tfrac{1}{2}\|v\|^2_{\mathcal{H}} + C\sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y_i\left(\langle v, \Phi(x_i)\rangle + b\right) \geq 1 - \xi_i \qquad \text{for } i = 1, 2, \ldots, m,\end{aligned}$$

The dual to this problem is always a finite dimensional Quadratic Program:

$$\min_{\alpha \in [0,C]^m} \tfrac{1}{2}\|\alpha\|^2_K - 1^T\alpha \qquad \text{s.t.} \quad \sum_{i=1}^m \alpha_i y_i = 0, \tag{12}$$

where K is the *Kernel matrix* given by $K_{ij} := y_i y_j \langle \Phi(x_i), \Phi(x_j)\rangle$. Once the dual is solved, v is recovered via $v = \sum_i \alpha_i y_i \Phi(x_i)$.

The choice of kernel is an art-form; there are many options available, and different kernels perform better on different problems. To develop a disciplined approach in this setting, the *multiple kernel learning* framework has been proposed. In this framework, we suppose that we have a choice of M kernels functions Φ_i with corresponding kernels K_i. We can now consider the linear combination

$$K(w) = \sum_{i=1}^M w_i K_i,$$

where w are some weights. A natural question is to find the best weighted combination; such an approach has been proposed in [24]. Specifically, requiring w to be in the unit simplex yields the problem

$$\min_{w \in \Delta_1} \widehat{f}(w) := \left\{ \min_{\alpha \in [0,C]^m} \tfrac{1}{2}\|\alpha\|^2_{K(w)} - 1^T\alpha \qquad \text{s.t.} \quad \sum_{i=1}^m \alpha_i y_i = 0 \right\}.$$

This problem was solved in [24] by using variable projection, with the outer problem solved by prox-gradient. Notice that each iteration requires a complete solve of a Quadratic Program.

Our novel approach switches which variable is projected out. Analogously to the trimmed regression case, a naive approach is given below:

$$\min_{\alpha \in [0,C]^m} \widetilde{f}(\alpha) - 1^T\alpha \qquad \text{s.t.} \qquad \sum_{i=1}^m \alpha_i y_i = 0, \tag{13}$$

where

$$\widetilde{f}(\alpha) = \min_{w\in\Delta_1} \tfrac{1}{2}\|\alpha\|^2_{K(w)} = \min_{w\in\Delta_1} \frac{1}{2}\sum_i w_i\langle K_i\alpha,\alpha\rangle \tag{14}$$

As in trimmed regression, this approach does not work, as $\widetilde{f}$ can be nonsmooth. We use the same smoothing modification, considering a modified function

$$\begin{aligned}\widetilde{f}_\beta(\alpha) &= \min_{w\in\Delta_1} \frac{1}{2}\sum_i w_i\langle K_i\alpha,\alpha\rangle + \frac{\beta}{2}\|w\|^2 \\ &= \frac{\beta}{2}\mathrm{dist}^2_{\Delta_k}\left(-\beta^{-1}g(\alpha)\right) - (2\beta)^{-1}\|g(\alpha)\|^2\end{aligned} \tag{15}$$

where $g_i(\alpha) = \alpha^T K_i\alpha$. Now, $\widetilde{f}_\beta(\alpha)$ is smooth, with $\nabla\widetilde{f}_\beta(\alpha) = \nabla g(\alpha)\bar{\alpha}$, where $\bar{\alpha}$ is the nearest point to $-\beta^{-1}g(\alpha)$ in the unit simplex. This projection has also been studied and implemented [9]. With the smoothing extension, standard approaches can be applied to solve (13) with the modified $\widetilde{f}_\beta$. In our numerical experiments, demonstrate the efficiency of this novel formulation.

3 Theory

In this section, we discuss some of the claimed derivative formulas in the aforementioned applications. To this end, recall that a function $f\colon \mathbb{R}^n \to \mathbb{R}$ is *differentiable* at $\bar{x}$ if there exists a vector v, called the *gradient* and denoted $\nabla f(\bar{x})$, satisfying

$$\lim_{y\to\bar{x}} \frac{f(y) - f(\bar{x}) - \langle v, y-\bar{x}\rangle}{\|y-\bar{x}\|} = 0.$$

On the other hand, a function f that is differentiable at $\bar{x}$ is *strictly differentiable* at $\bar{x}$ if it satisfies the slightly stronger property

$$\lim_{z,x\to\bar{x}} \frac{f(z) - f(x) - \langle \nabla f(\bar{x}), z-x\rangle}{\|z-x\|} = 0.$$

The following fundamental result [25][Theorem 10.58] is key for establishing derivative formulas for $\widetilde{g}$ in (8) and for $\widetilde{f}$ in (9). Strict differentiability, in particular, implies local Lipschitz continuity of the function, in contrast to conventional differentiability [25, Theorem 9.18].

Theorem 1 (Derivative of the projected function)**.** *Consider a function $f\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ satisfying the following properties*

1. *f is continuous;*
2. *$f(x,z)$ is level-bounded in z locally uniformly in x, meaning that for any $\alpha\in\mathbb{R}$ and any compact set $X\subset\mathbb{R}^n$, the union of sublevel sets $\bigcup_{x\in X}\{z : f(x,z)\le\alpha\}$ is bounded.*
3. *gradient $\nabla_x f(x,z)$ exists for all (x,z) and depends continuously on (x,z).*

Define now the projected function and the minimizing set:

$$\widetilde{f}(x) := \inf_z f(x,z) \qquad \text{and} \qquad P(x) := \operatorname*{argmin}_z f(x,z).$$

Then $\widetilde{f}$ is strictly differentiable at any point x for which the set $Y(x) := \{\nabla_x f(x,z) : z\in P(x)\}$ is a singleton, and in this case equality $\nabla\widetilde{f}(x) = Y(x)$ holds.

Let us see how this theorem applies. Consider the function $\widetilde{g}$ in the formulation (8). Clearly the inner function satisfies properties 1 and 3 of Theorem 1. Property 2 follows quickly from strong convexity of p and the assumptions that g is lower bounded and $H(x)$ is invertible. Then the set $Y(x)$ is clearly a singleton, and the claimed derivative formula follows (even if p is nonsmooth).

Next, consider the function $\widetilde{f}$ in (9). Again, properties 1 and 3 of Theorem 1 hold. To justify property 2, we must assume that $\Phi(\theta)$ has full column rank for all θ. Then property 2

follows immediately. Since the objective is strongly convex in a, the set $Y(\theta)$ is a a singleton and the derivative formula follows.

Next note that there are important limitations to Theorem 1. We already saw that the naive projected function $x \mapsto \min_{w \in \Delta_k \cap \mathbb{B}_\infty} \sum_i w_i g_i(x)$ in equation (10) can easily be nonsmooth; see figure 1. The same occurs for the naive formulation (14). The difficulty is that the parameter that is projected out varies over a constrained set, and hence properties 1 and 3 decisively fail.

As we saw in sections 2.3 and 2.4, we can circumvent these issues by smoothing the projected problem while still preserving the structure of the solution set $P(x)$ (feasibility, sparsity, etc.) This technique seems promising much more generally for problems of the form

$$\min_{x,z} \; g(z,x) := h(z, g(x)) + r(x),$$

where h, r, and g satisfy appropriate conditions. Such a smoothing technique in nonsmooth optimization appears to be new, and will be the subject of further investigation.

4 Numerical Illustrations

Direction-of-arrival estimation (DOA)

A classical problem in array-based signal processing is the estimation of the *direction-of-arrival*. In a typical setup, incoming signals from multiple sources (e.g., electromagnetic or acoustic waves) are recorded by an array of receivers. The goal is to identify the directions from which various contributions originated. This problem falls in the class of exponential data-fitting, as an incoming plane wave can be described by a complex exponential. The signal model is

$$y = \Phi(\theta)a + \epsilon,$$

where $y \in \mathbb{R}^m$ denotes the signal, $\Phi(\theta) \in \mathbb{R}^{m \times n}$ is a matrix with elements $\exp(-\imath\theta_j \cdot x_i)$, $a \in \mathbb{R}^n$ are the amplitudes of the individual components; $\theta_j \in \mathbb{S}^1$ are the directions and $x_i \in \mathbb{R}^2$ are the locations of the receivers.

The classical way of estimating the parameters θ_i is the MUSIC algorithm [8, 30]. This algorithm proceeds to estimate the DOA as follows. First, we treat both a and ϵ as random variables and define the signal correlation matrix

$$\Sigma_\theta = \mathbb{E}_a\{\Phi(\theta)aa^*\Phi(\theta)^*\},$$

which is an $m \times m$ matrix with rank $\min\{m, n\}$. We let P be the matrix whose columns span the null-space of Σ_θ. The MUSIC pseudo-spectrum is now defined as

$$f(\theta) = \frac{1}{\|P^* s(\theta)\|_2},$$

where $s_i(\theta) = \exp(\imath\theta \cdot x_i)$. If $s(\theta)$ is nearly in the range of Σ_θ, its projection onto the null-space is nearly zero, thus causing a large peak in the pseudo-spectrum. The components in the signal can then be detected by taking the m largest peaks in the pseudo-spectrum.

In practice, we do not have access to the signal correlation. Instead, we construct the sample average

$$\Sigma_y \approx \frac{1}{N} \sum_{i=1}^{N} y_i y_i^*.$$

If the ϵ is i.i.d. Gaussian with variance σ and independent of a, we have

$$\Sigma_y = \Sigma_\theta + \sigma^2 I,$$

in which case we construct P from the singular vectors of Σ_y with singular values $< \sigma^2$. Proper identification of the null-space requires an estimate of the noise level.

MUSIC vs. Nonsmooth DOA (full and reduced) An example with $N = 10$, $n = 101$ and $m = 5$ is shown in figure 2. Note that the observed data contains arrivals from only 3 distinct directions. We solved the LASSO problems using an interior point method (IPsolve [3]) while using a Quasi-Newton method to solve the outer optimization problem. We compare this variable projection approach to solving the full optimization problem in (a, θ) jointly using a non-smooth Quasi-Newton method [16].

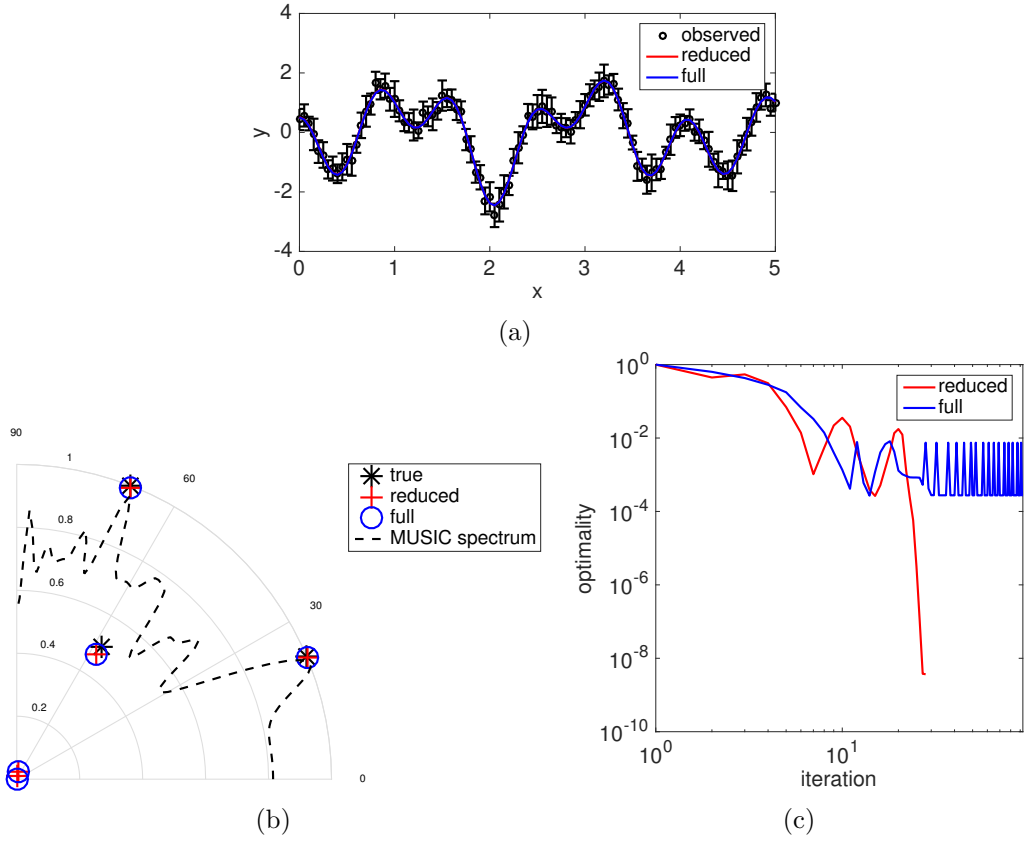


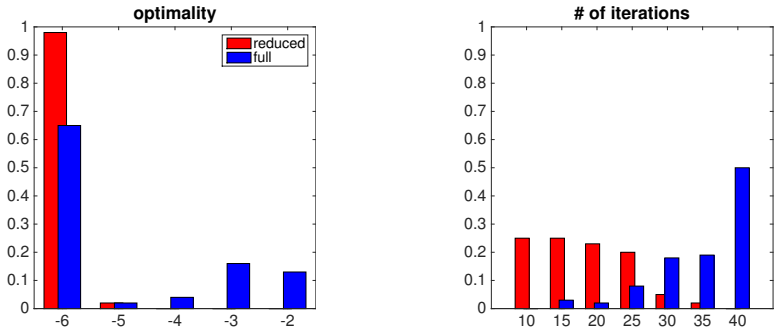
Figure 2: (a) observed and fitted data, (b) true and estimated direction-of-arrival and corresponding amplitudes, as well as the MUSIC spectrum, and (c) convergence of the full and reduced approach.

The observed and fitted data are shown in figure 2 (a) and the true and estimated DOAs and amplitudes as well as the MUSIC spectrum are shown in figure 2 (b). The MUSIC spectrum clearly picks up the strongest modes but fails to recover the weaker mode. Both the full and variable projections approaches correctly estimate both the directions and amplitudes of the 3 arrivals. The full approach, however, fails to converge as can be seen in 2 (c).

We repeated the above experiment for 100 different initial guesses and summarized the results in terms of the number of iterations required to reach the desired tolerance of 10^{-6} and optimality (relative norm of the gradient) in figure 3. We see that the reduced approach almost always reaches the desired tolerance and requires less iterations to do so. In contrast, the full approach fails to reach the desired tolerance in roughly 35/100 cases.

Blind sparse deconvolution: full vs. reduced In this example, we discuss a parametrized blind deconvolution problem. We are given a noisy image which we want to decompose into a number of parametrized basis functions. Such problems occur frequently in astrophysics [33], image deblurring [2, 20] and seismology [17, 26].

In this case we we let $\varphi_{ij}(r, s) = \exp(-s_j^2 \|x_i - r_j\|_2^2)$ and use the interior point method IPsolve [3] to solve the LASSO problems with positivity constraints. The outer optimization over (r, s) is done using a Quasi-Newton method. We compare the VP approach to a full optimization over the (a, r, s)-space. The results are shown in figure 4. The ground truth image consists of four blobs; the initial guess for r is obtained by picking the 50 largest peaks



(a) (b)

Figure 3: Results of reduced versus full-space optimization for 100 different initial guesses, with a required tolerance of 10^{-6}. (a) shows the relative tolerance upon convergence (or failure) of the optimization, (b) shows the number of iterations required by both methods. We see that the reduced approach almost always reaches the desired tolerance and requires less iterations than the full approach.

in the noisy image. The reconstructions obtained by the reduced and full approaches look similar, but we see that the full approach has trouble getting rid of the superfluous points. The convergence plots, shown in figure 5, show that the full approach does not fit the data as well as the reduced approach and converges very slowly.

Robust logistic regression using variable projection Logistic regression is a popular alternative to support vector machines that seeks to find the best hyperplane to separate two classes of points. It can be formulated as an optimization problem as follows:

$$\min_{\theta} f(\theta) := \sum_{i} \log(1 + \exp(-y_i x_i^T \theta)) + R(\theta), \tag{16}$$

where x_i represents the ith feature, $y_i \in \{-1, +1\}$ is a class label, θ describes the hyperplane we seek, and $R(\theta)$ is a regularizer; in this example, we use a small multiple of $\|\cdot\|^2$ as the regularizer. In this section, we consider a classification problem where a portion of the data has been contaminated. We pick 2000 features in $\mathbb{R}^{200}$, and we replace 10% (200 features) with random noise, with entries on average ten times larger than those of actual features. The smoothed trimmed approach for this problem is given by

$$\min_{\theta} \widetilde{f}(\theta) = \min_{w \in \Delta_k \cap \mathbb{B}_\infty} \left\{ \sum_{i} w_i \log(1 + \exp(-y_i x_i^T \theta)) + \frac{\beta}{2}\|w\|^2 \right\} + R(\theta), \tag{17}$$

as described in section 2. We then compare three scenarios:

1. Data has not been contaminated, standard logistic regression used
2. Data has been contaminated, standard logistic regression used
3. Data has been contaminated, robust logistic regression (17) used with $k = 1000$.

Note that the algorithm doesn't use information about the exact number of contaminated features. Rather, it assumes that at least half of the data is good. The results were nearly identical without using regularization $R(\theta)$. We used $\beta = 1$ for the smoothing.

Fast MKL using variable projection In this example, we show how variable projection can be used for Multiple Kernel Learning. We solve the smoothed dual problem (15) using SQP with approximate Hessian $H = \sum_i \overline{w_i} K_i$. We consider two parametrized classes of kernel functions, polynomial and Gaussian:

$$k_p^{\text{pol}}(x, x') = \left(1 + x^T x'\right)^p,$$

$$k_s^{\text{Gauss}}(x, x') = \exp(-\|x - x'\|_2^2 / a^2).$$

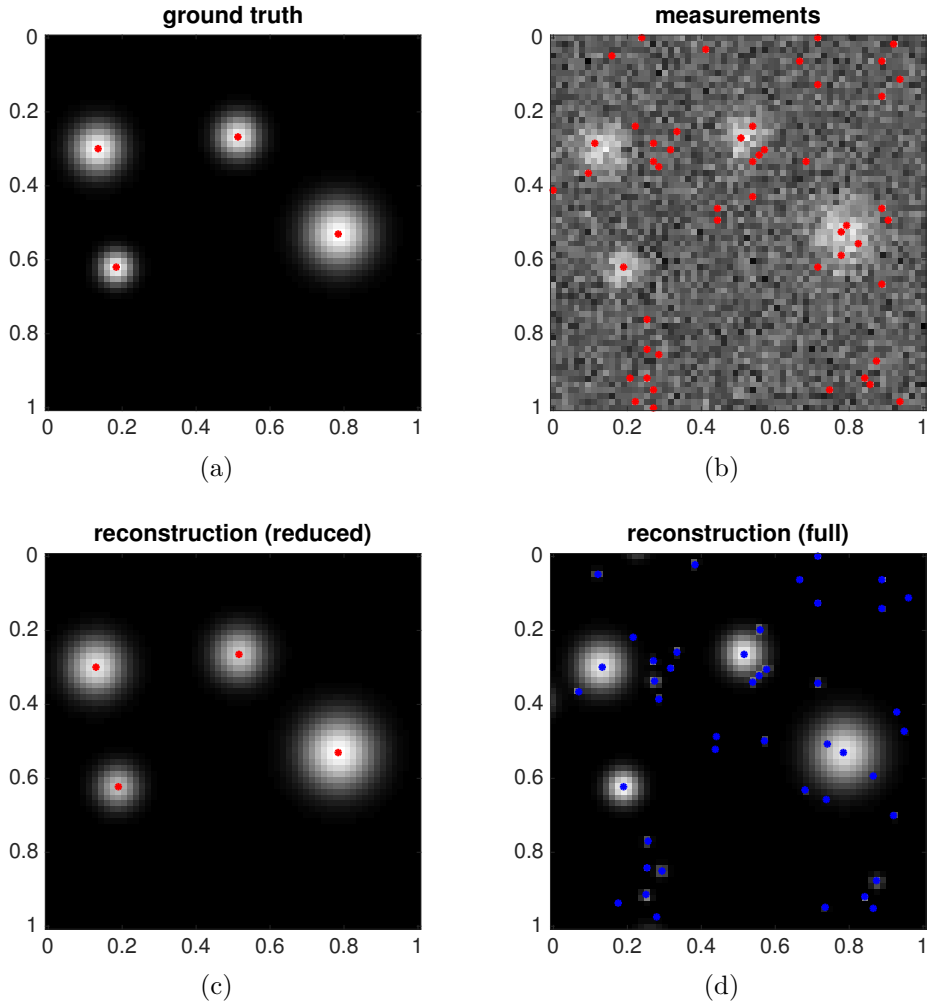


Figure 4: (a) ground truth image, (b) noisy image and initial guess for c, (c) results using the reduced approach, and (d) results using the full approach.

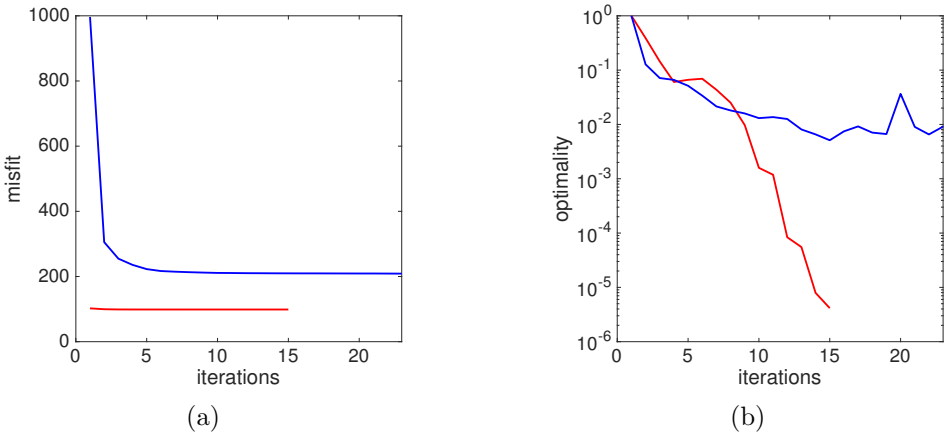


Figure 5: Convergence of the full and reduced approaches in terms of the misfit (a) and the optimality (b).

As an illustrative example, we consider classifying points in $\mathbb{R}^2$ to the left and to the right of an elliptic curve. In figure 6, we show the results for various kernels. In table 3, we show the number of iterations required for each kernel.

In figure 7 we show the result using a total of 12 kernels; 5 polynomial kernels with

	Fraction Correct	Standard Deviation
No contamination	0.95	0.006
Contam. and standard LR	0.74	0.02
Contam. and robust LR	0.86	0.02

Table 2: Results averaged over 20 runs for three scenarios: no contamination, contamination with standard logistic regression, and contamination with smoothed trimmed approach. The test set consists of 1000 data points. "Standard Deviation" of the "Fraction Correct" is taken over 20 experiments.

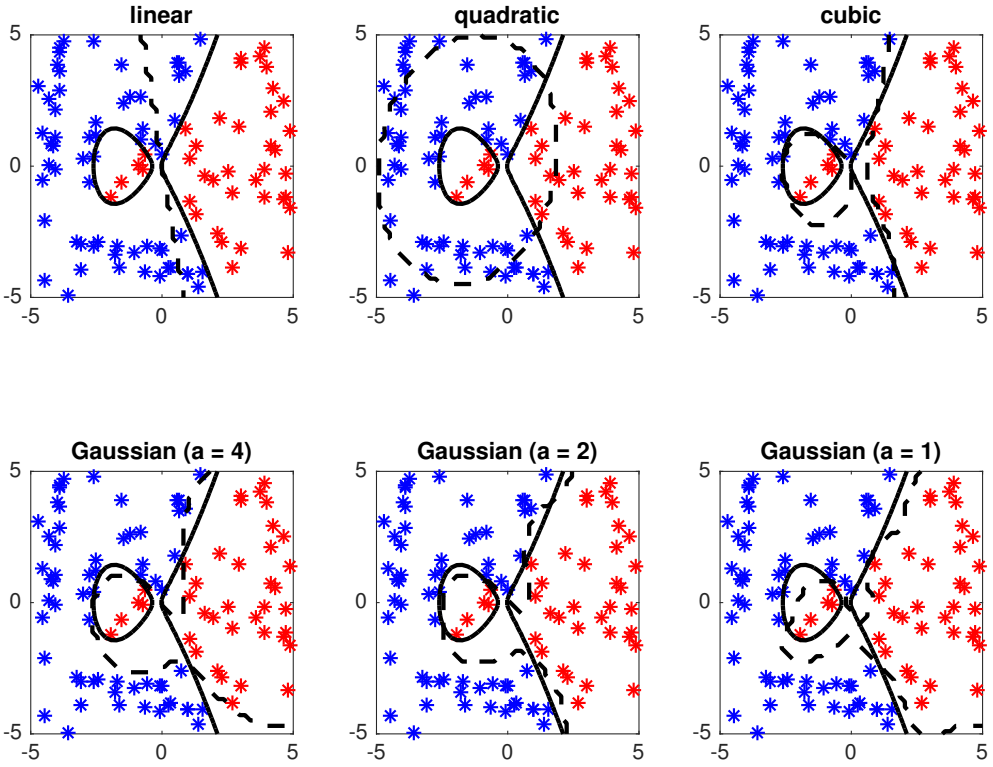


Figure 6: Classification of points in the plane using different kernels. The red and blue stars indicate the training set, the solid line denotes the elliptic curve used to separate the points and the dotted line shows the classification resulting from the optimization procedure.

$p = 1, 2, \ldots, 5$ and 7 Gaussian kernels with parameters $a = 1, 1.5, 2, \ldots, 4$. Our approach quickly hones in on the Gaussian kernel with $a = 1$ and takes a total of 107 iterations.

kernel	1	2	3	4	5	6
iterations	30	33	21	94	57	45

Table 3: Number of iterations required for the results shown in figure 6.

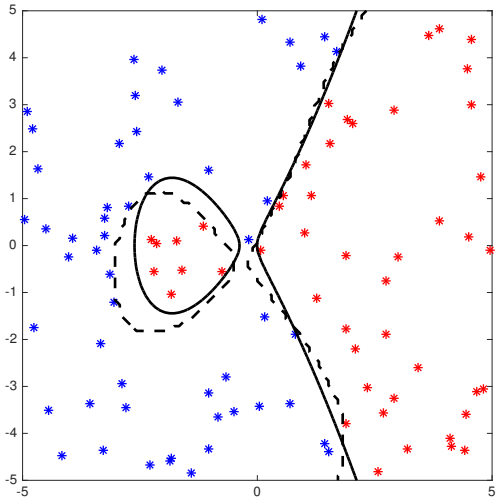


Figure 7: Classification of points in the plane using the new multiple kernel learning approach.

5 Conclusions

Variable projection has been successfully used in a variety of contexts; the popularity of the approach is largely due to its superior numerical performance when compared to joint optimization schemes. In this paper, we considered a range of nonsmooth applications, illustrating the use of variable projection for sparse deconvolution and direction of arrivals estimation. We showed that differentiability of the projected function can be understood using basic variational analysis, and that this approach has limitations and can fail for interesting cases. In particular, for robust formulations and multiple kernel learning, we showed that the projected function can fail to be differentiable. To circumvent this difficulty, we proposed a novel smoothing technique for the inner problem that preserves feasibility and structure of the solution set, while guaranteeing differentiability of the projected function. Numerical examples in all cases showed that variable projection schemes, when appropriately applied, are highly competitive for a wide range of nonsmooth applications.

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] G. S. Abdoulaev, K. Ren, and A. H. Hielscher. Optical tomography as a PDE-constrained optimization problem. *Inverse Problems*, 21(5):1507–1530, oct 2005.

[2] M. S. C. Almeida and M. a. T. Figueiredo. Parameter estimation for blind and non-blind deblurring using residual whiteness measures. *IEEE Transactions on Image Processing*, 22(7):2751–2763, 2013.

[3] A. Y. 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.

[4] A. Y. Aravkin and T. van Leeuwen. Estimating nuisance parameters in inverse problems. *Inverse Problems*, 28(11):115016, nov 2012.

[5] B. Bell and J. Burke. Algorithmic differentiation of implicit functions and optimal values. In C. H. Bischof, H. M. Bücker, P. D. Hovland, U. Naumann, and J. Utke, editors, *Advances in Automatic Differentiation*, pages 67–77. Springer, 2008.

[6] G. Biros and O. Ghattas. Inexactness issues in the Lagrange-Newton-Krylov-Schur method for PDE-constrained optimization. In *Large-Scale PDE-Constrained Optimization*, pages 93–114. Springer, 2003.

[7] J. Borwein and A. Lewis. *Convex analysis and nonlinear optimization*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 3. Springer, New York, second edition, 2006. Theory and examples.

[8] M. Cheney. The linear sampling method and the MUSIC algorithm. *Inverse Problems*, 17(4):591–595, 2001.

[9] L. Condat. Fast Projection onto the Simplex and the l1 Ball. Technical report, 2014.

[10] J. Dennis, M. Heinkenschloss, and L. Vicente. Trust-region interior-point SQP algorithms for a class of nonlinear programming problems. *SIAM Journal on Control and Optimization*, 36(5):1750–1794, 1998.

[11] G. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares which variables separate. *SIAM J. Numer. Anal.*, 10(2):413–432, 1973.

[12] G. Golub and V. Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Inverse Problems*, 19(2):R1, 2003.

[13] E. Haber, U. M. Ascher, and D. Oldenburg. On optimization techniques for solving nonlinear inverse problems. *Inverse Problems*, 16(5):1263–1280, oct 2000.

[14] S. M. Hanasoge. Full waveform inversion of solar interior flows. *The Astrophysical Journal*, 797(1):23, nov 2014.

[15] R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

[16] A. S. Lewis and M. L. Overton. Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, 141(1-2):135–163, oct 2013.

[17] T. T. Y. Lin and F. J. Herrmann. Robust estimation of primaries by sparse inversion via one-norm minimization. *Geophysics*, 78(3):R133–R150, 2013.

[18] M. Osborne. Separable least squares, variable projection, and the Gauss-Newton algorithm. *Electronic Transactions on Numerical Analysis*, 28(2):1–15, 2007.

[19] D. P. OLeary and B. W. Rust. Variable projection for nonlinear least squares problems. *Computational Optimization and Applications*, 54(3):579–593, apr 2013.

[20] S. U. Park, N. Dobigeon, and A. O. Hero. Semi-blind sparse image reconstruction with application to MRFM. *IEEE Transactions on Image Processing*, 21(9):3838–3849, 2012.

[21] V. Pereyra and G. Scherer, editors. *Exponential Data Fitting and its Applications*. Bentham Science and Science Publishers, mar 2012.

[22] P. Philip. Optimal Control and Partial Differential Equations. Technical report, 2009.

[23] R.-E. Plessix. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, 167(2):495–503, 2006.

[24] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. More efficiency in multiple kernel learning. In *Proceedings of the 24th international conference on Machine learning*, pages 775–782. ACM, 2007.

[25] R. Rockafellar and R. Wets. *Variational Analysis*, volume 317. Springer, 1998.

[26] A. A. Royer, M. G. Bostock, and E. Haber. Blind deconvolution of seismograms regularized via minimum support. *Inverse Problems*, 28(12):125010, dec 2012.

[27] A. Ruhe and P. Wedin. Algorithms for separable nonlinear least squares problems. *SIAM Rev.*, 22(3):318–337, 1980.

[28] D. Ruppert and R. J. Carroll. Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, 75(372):828–838, 1980.

[29] S. Saitoh. *Theory of reproducing kernels and its applications*, volume 189 of *Pitman Research Notes in Mathematics Series*. Longman Scientific & Technical, Harlow; copublished in the United States with John Wiley & Sons, Inc., New York, 1988.

[30] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, mar 1986.

[31] B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In *Computational learning theory (Amsterdam, 2001)*, volume 2111 of *Lecture Notes in Comput. Sci.*, pages 416–426. Springer, Berlin, 2001.

[32] B. Schölkopf and A. J. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[33] P. Shearer and A. C. Gilbert. A generalization of variable elimination for separable inverse problems beyond least squares. *Inverse Problems*, 29(4):045003, Apr. 2013.

[34] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[35] T. van Leeuwen and F. J. Herrmann. A penalty method for PDE-constrained optimization in inverse problems. Technical Report 1, jan 2016.

[36] W. Wang and M. A. Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*, 2013.

[37] E. Yang and A. C. Lozano. Robust gaussian graphical modeling with the trimmed graphical lasso. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2584–2592. Curran Associates, Inc., 2015.