Optimization Methods for Inverse Problems

Nan Ye, Farbod Roosta-Khorasani, Tiangang Cui

Abstract Optimization plays an important role in solving many inverse problems. Indeed, the task of inversion often either involves or is fully cast as a solution of an optimization problem. In this light, the mere non-linear, non-convex, and large-scale nature of many of these inversions gives rise to some very challenging optimization problems. The inverse problem community has long been developing various techniques for solving such optimization tasks. However, other, seemingly disjoint communities, such as that of machine learning, have developed, almost in parallel, interesting alternative methods which might have stayed under the radar of the inverse problem community. In this survey, we aim to change that. In doing so, we first discuss current state-of-the-art optimization methods widely used in inverse problems. We then survey recent related advances in addressing similar challenges in problems faced by the machine learning community, and discuss their potential advantages for solving inverse problems. By highlighting the similarities among the optimization challenges faced by the inverse problem and the machine learning communities, we hope that this survey can serve as a bridge in bringing together these two communities and encourage cross fertilization of ideas.

1 Introduction

Inverse problems arise in many applications in science and engineering. The term "inverse problem" is generally understood as the problem of finding a specific physical property, or properties, of the medium under investigation, using indirect measurements. This is a highly important field of applied mathematics and scientific

Nan Ye
ACEMS & Queensland University of Technology, e-mail: n.ye@qut.edu.au

Farbod Roosta-Khorasani
University of Queensland e-mail: fred.roostauq.edu.au

Tiangang Cui
Monash University e-mail: tiangang.cui@monash.edu

computing, as to a great extent, it forms the backbone of modern science and engineering. Examples of inverse problems can be found in various fields within medical imaging [6, 7, 12, 74, 98] and several areas of geophysics including mineral and oil exploration [8, 18, 77, 99].

In general, an inverse problem aims at recovering the unknown underlying parameters of a physical system which produces the available observations/measurements. Such problems are generally ill-posed [55]. This is often solved via two approaches: a Bayesian approach which computes a posterior distribution of the models given prior knowledge and the data, or a regularized data fitting approach which chooses an optimal model by minimizing an objective that takes into account both fitness to data and prior knowledge. The Bayesian approach can be used for a variety of downstream inference tasks, such as credible intervals for the parameters; it is generally more computationally expensive than the data fitting approach. The computational attractiveness of data fitting comes at a cost: it can only produce a "point" estimate of the unknown parameters. However, in many applications, such a point estimate can be more than adequate.

In this review, we focus on the data fitting approach. The approach consists of the four building blocks: a parametric model of the underlying physical phenomenon, a forward solver that predicts the observation given the model parameters, an objective function measuring how well a model fits the observation, and an optimization algorithm for finding model parameters optimizing the objective function. The first three components together conceptually defines what an optimal model is, and the optimization algorithm provides a computational means to find the optimal model (usually requires solving the forward problem during optimization). Each of these four building blocks is an active area of research. This paper focuses on the optimization algorithms. While numerous works have been done on the subject, there are still many challenges remaining, including scaling up to large-scale problems, dealing with non-convexity. On the other hand, optimization also constitutes a backbone of machine learning [17, 35]. Consequently, there are many related developments in optimization from the machine learning community. However, thus far and rather independently, the machine learning and the inverse problems communities have largely developed their own sets of tools and algorithms to address their respective optimization challenges. It only stands to reason that many of the recent advances by machine learning can be potentially applicable for addressing challenges in solving inverse problems. We aim to bring out this connection and encourage permeation of ideas across these two communities.

In Section 2, we present general formulations for the inverse problem, some typical inverse problems, and optimization algorithms commonly used to solve the data fitting problem. We discuss recent advances in optimization in Section 3. We then discuss areas in which cross-fertilization of optimization and inverse problems can be beneficial in Section 4. We conclude in Section 5. We remark that our review of these recent developments focus on iterative algorithms using gradient and/or Hessian information to update current solution. We do not examine global optimization methods, such as genetic algorithms, simulated annealing, particle swarm optimization, which have also received increasing attention recently (e.g. see [101]).

2 Inverse Problems

An inverse problem can be seen as the reverse process of a forward problem, which concerns with predicting the outcome of some measurements given a complete description of a physical system. Mathematically, a physical system is often specified using a set of model parameters $\mathbf{m}$ whose values completely characterize the system. The model space $\mathscr{M}$ is the set of possible values of $\mathbf{m}$. While $\mathbf{m}$ usually arises as a function, in practice it is often discretized as a parameter vector for the ease of computation, typically using the finite element method, the finite volume method, or the finite difference method. The forward problem can be denoted as

$$\mathbf{m} \to \mathbf{d} = \mathbf{f}(\mathbf{m}), \tag{1}$$

where $\mathbf{d}$ are the error-free predictions, and the above notation is a shorthand for $\mathbf{d} = (\mathbf{d}_1, \ldots, \mathbf{d}_s) = (\mathbf{f}_1(\mathbf{m}), \ldots, \mathbf{f}_s(\mathbf{m}))$, with $\mathbf{d}_i \in \mathbb{R}^l$ being the i-th measurement. The function $\mathbf{f}$ represents the physical theory used for the prediction and is called the forward operator. The observed outcomes contain noises and relate to the system via the following the observation equation

$$\mathbf{d} = \mathbf{f}(\mathbf{m}) + \boldsymbol{\eta}, \tag{2}$$

where $\boldsymbol{\eta}$ are the noises occurred in the measurements. The inverse problem aims to recover the model parameters $\mathbf{m}$ from such noisy measurements.

The inverse problem is almost always ill-posed, because the same measurements can often be predicted by different models. There are two main approaches to deal with this issue. The Bayesian approach assumes a prior distribution $P(\mathbf{m})$ on the model and a conditional distribution $P(\boldsymbol{\eta} \mid \mathbf{m})$ on noise given the model. The latter is equivalent to a conditional distribution $P(\mathbf{d} \mid \mathbf{m})$ on measurements given the model. Given some measurements $\mathbf{d}$, a posterior distribution $P(\mathbf{m} \mid \mathbf{d})$ on the models is then computed using the Bayes rule

$$P(\mathbf{m} \mid \mathbf{d}) \propto P(\mathbf{m}) P(\mathbf{d} \mid \mathbf{m}). \tag{3}$$

Another approach sees the inverse problem as a data fitting problem that finds an parameter vector $\mathbf{m}$ that gives predictions $\mathbf{f}(\mathbf{m})$ that best fit the observed outcomes $\mathbf{d}$ in some sense. This is often cast as an optimization problem

$$\min_{\mathbf{m} \in \mathscr{M}} \quad \psi(\mathbf{m}, \mathbf{d}), \tag{4}$$

where the misfit function ψ measures how well the model $\mathbf{m}$ fits the data $\mathbf{d}$. When there is a probabilistic model of $\mathbf{d}$ given $\mathbf{m}$, a typical choice of $\psi(\mathbf{m}, \mathbf{d})$ is the negative log-likelihood. Regularization is often used to address the issue of multiple solutions, and additionally has the benefit of stabilizing the solution, that is, the solution is less likely to change significantly in the presence of outliers [5, 39, 111]. Regularization incorporats some *a priori* information on $\mathbf{m}$ in the form of a regularizer $R(\mathbf{m})$ and solves the regularized optimization problem

$$\min_{\mathbf{m}\in\mathscr{M}} \quad \psi_{R,\alpha}(\mathbf{m},\mathbf{d}) := \psi(\mathbf{m},\mathbf{d}) + \alpha R(\mathbf{m}), \tag{5}$$

where $\alpha > 0$ is a constant that controls the tradeoff between prior knowledge and the fitness to data. The regularizer $R(\mathbf{m})$ encodes a preference over the models, with preferred models having smaller R values. The formulation in Eq. (5) can often be given a *maximum a posteriori (MAP)* interpretation within the Bayesian framework [100]. Implicit regularization also exists in which there is no explicit term $R(\mathbf{m})$ in the objective [30, 32, 56, 57, 89, 90].

The misfit function often has the form $\phi(\mathbf{f}(\mathbf{m}),\mathbf{d})$, which measures the difference between the prediction $\mathbf{f}(\mathbf{m})$ and the observation $\mathbf{d}$. For example, ϕ may be chosen to be the Euclidean distance between $\mathbf{f}(\mathbf{m})$ and $\mathbf{d}$. In this case, the regularized problem takes the form

$$\min_{\mathbf{m}\in\mathscr{M}} \quad \phi_{R,\alpha}(\mathbf{m},\mathbf{d}) := \phi(\mathbf{f}(\mathbf{m}),\mathbf{d}) + \alpha R(\mathbf{m}), \tag{6}$$

This can also be equivalently formulated as choosing the most preferred model satisfying constraints on its predictions

$$\min_{\mathbf{m}\in\mathscr{M}} \quad R(\mathbf{m}), \quad \text{s.t.} \quad \phi(\mathbf{f}(\mathbf{m}),\mathbf{d}) \le \rho. \tag{7}$$

The constant ρ usually relates to noise and the maximum discrepancy between the measured and the predicted data, and can be more intuitive than α.

2.1 PDE-Contrained Inverse Problems

For many inverse problems in science and engineering, the forward model is not given explicitly via a forward operator $\mathbf{f}(\mathbf{m})$, but often conveniently specified via a set of partial differential equations (PDEs). For such problems, Eq. (6) has the form

$$\min_{\mathbf{m}\in\mathscr{M},\mathbf{u}} \quad \phi(P\cdot\mathbf{u},\mathbf{d}) + \alpha R(\mathbf{m}), \quad \text{s.t.} \quad c_i(\mathbf{m},\mathbf{u}_i) = 0, \quad i = 1,\dots,s, \tag{8}$$

where $P\cdot\mathbf{u} = (P_1,\dots,P_s)\cdot(\mathbf{u}_1,\dots,\mathbf{u}_s) = (P_1\mathbf{u}_1,\dots,P_s\mathbf{u}_s)$ with $\mathbf{u}_i$ being the field in the i-th experiment, P_i being the projection operator that selects fields at measurement locations in $\mathbf{d}_i$ (that is, $P_i\mathbf{u}_i$ are the predicted values at locations measured in $\mathbf{d}_i$), and $c_i(\mathbf{m},\mathbf{u}_i) = 0$ corresponds to the forward model in the i-th experiment. In practice, the forward model can often be written as

$$\mathscr{L}_i(\mathbf{m})\mathbf{u}_i = \mathbf{q}_i, \quad i = 1,\dots,s, \tag{9}$$

where $\mathscr{L}_i(\mathbf{m})$ is a differential operator, and $\mathbf{q}_i$ is a term that incorporates source terms and boundary values.

The fields $\mathbf{u}_1,\dots,\mathbf{u}_s$ in Eq. (8) and Eq. (9) are generally functions in two or three dimensional spaces, and finding closed-form solutions is usually not possible. Instead, the PDE-constrained inverse problem is often solved numerically by

discretizing Eq. (8) and Eq. (9) using the finite element method, the finite volume method, or the finite difference method. Often the discretized PDE-constrained inverse problem takes the form

$$\min_{\mathbf{m}\in\mathscr{M},\mathbf{u}} \quad \phi(P\mathbf{u},\mathbf{d}) + \alpha R(\mathbf{m}), \quad \text{s.t.} \quad L_i(\mathbf{m})\mathbf{u}_i = \mathbf{q}_i, \quad i = 1,\ldots,s, \tag{10}$$

where P is a block-diagonal matrix consisting of diagonal blocks $P_1,\ldots,P_s$ representing the discretized projection operators, $\mathbf{u}$ is the concatenation of the vectors $\mathbf{u}_1,\ldots,\mathbf{u}_s$ representing the discretized fields, and each $L_i(\mathbf{m})$ is a square, nonsingular matrix representing the differential operator $\mathscr{L}_i(\mathbf{m})$. Each $L_i(\mathbf{m})$ is typically large and sparse. We abuse the notations P, $\mathbf{u}$ to represent both functions and their discretized versions, but the meanings of these notations will be clear from context.

The constrained problem in Eq. (10) can be written in an unconstrained form by eliminating $\mathbf{u}$ using $\mathbf{u}_i = L_i^{-1}\mathbf{q}_i$,

$$\min_{\mathbf{m}\in\mathscr{M}} \quad \phi(PL^{-1}(\mathbf{m})\mathbf{q},\mathbf{d}) + \alpha R(\mathbf{m}), \tag{11}$$

where L is the block-diagonal matrix with $L_1,\ldots,L_s$ as the diagonal blocks, and $\mathbf{q}$ is the concatenation of $\mathbf{q}_1,\ldots,\mathbf{q}_s$. Note that, as in the case of (6), here we have $\mathbf{f}(\mathbf{m}) = PL^{-1}(\mathbf{m})\mathbf{q}$.

Both the constrained and unconstrained formulations are used in practice. The constrained formulation can be solved using the method of Lagrangian multipliers. This does not require explicitly solving the forward problem as in the unconstrained formulation. However, the problem size increases, and the problem becomes one of finding a saddle point of the Lagrangian, instead of finding a minimum as in the constrained formulation.

2.2 Image Reconstruction

Image reconstruction studies the creation of 2-D and 3-D images from sets of 1-D projections. The 1-D projections are generally line integrals of a function representing the image to be reconstructed. In the 2-D case, given an image function $f(x,y)$, the integral along the line at a distance of s away from the origin and having a normal which forms an angle ϕ with the x-axis is given by the Randon transform

$$p(s,\phi) = \int_{-\infty}^{\infty} f(z\sin\phi + s\cos\phi, -z\cos\phi + s\sin\phi)dz. \tag{12}$$

Reconstruction is often done via back projection, filtered back projection, or iterative methods [58, 80]. Back projection is the simplest but often results in a blurred reconstruction. Filtered back projection (FBP) is the analytical inversion of the Radon transform and generally yields reconstructions of much better quality than back projection. However, FBP may be infeasible in the presence of discontinuities

or noise. Iterative methods take noise into account, by assuming a distribution for the noise. The objective function is often chosen to be a regularized likelihood of the observation, which is then iteratively optimized using the expectation maximization (EM) algorithm.

2.3 Objective Function

One of the most commonly used objective function is the least squares criterion, which uses a quadratic loss and a quadratic regularizer. Assume that the noise for each experiment in (2) is independently but normally distributed, i.e., $\boldsymbol{\eta}_i \sim \mathcal{N}(0, \Sigma_i), \forall i$, where $\Sigma_i \in \mathbb{R}^{l\times l}$ is the covariance matrix. Let Σ be the block-diagonal matrix with $\Sigma_1, \ldots, \Sigma_s$ as the diagonal blocks. The standard *maximum likelihood* (ML) approach [100], leads to minimizing the least squares (LS) misfit function

$$\phi(\mathbf{m}) := \|\mathbf{f}(\mathbf{m}) - \mathbf{d}\|^2_{\Sigma^{-1}}, \tag{13}$$

where the norm $\|x\|_A = \sqrt{x^\top A x}$ is a generalization of the Euclidean norm (assuming the matrix A is positive definite, which is true in the case of Σ_i^{-1}). In the above equation, we simply write the general misfit function $\phi(\mathbf{f}(\mathbf{m}), \mathbf{d})$ as $\phi(\mathbf{m})$ by taking the measurements $\mathbf{d}$ as fixed and omitting it from the notation. As previously discussed, we often minimize a regularized misfit function

$$\phi_{R,\alpha}(\mathbf{m}) := \phi(\mathbf{m}) + \alpha R(\mathbf{m}). \tag{14}$$

The prior $R(\mathbf{m})$ is often chosen as a Gaussian regularizer $R(\mathbf{m}) = (\mathbf{m} - \mathbf{m}_{\text{prior}})^\top \Sigma_m^{-1} (\mathbf{m} - \mathbf{m}_{\text{prior}})$. We can also write the above optimization problem as minimizing $R(\mathbf{m})$ under the constraints

$$\sum_{i=1}^{s} \|\mathbf{f}_i(\mathbf{m}) - \mathbf{d}_i\| \le \rho. \tag{15}$$

The least-squares criterion belongs to the class of ℓ_p-norm criteria, which contain two other commonly used criteria: the least-absolute-values criterion and the minimax criterion [107]. These correspond to the use of the ℓ_1-norm and the ℓ_∞-norm for the misfit function, while the least squares criterion uses the ℓ_2-norm. Specifically, the least-absolute-values criterion takes $\phi(\mathbf{m}) := \|\mathbf{f}(\mathbf{m}) - \mathbf{d}\|_1$, and the minimax criterion takes $\phi(\mathbf{m}) := \|\mathbf{f}(\mathbf{m}) - \mathbf{d}\|_\infty$. More generally, each coordinate in the difference may be weighted. The ℓ_1 solution is more robust (that is, less sensitive to outliers) than the ℓ_2 solution, which is in turn more robust than the ℓ_∞ solution [25]. The ℓ_∞ norm is desirable when outliers are uncommon but the data are corrupted by uniform noise such as the quantization errors [26].

Besides the ℓ_2 regularizer discussed above, the ℓ_1-norm is often used too. The ℓ_1 regularizer induces sparsity in the model parameters, that is, heavier ℓ_1 regularization leads to fewer non-zero model parameters.

2.4 Optimization Algorithms

Various optimization techniques can be used to solve the regularized data fitting problem. We focus on iterative algorithms for nonlinear optimization below as the objective functions are generally nonlinear. In some cases, the optimization problem can be transformed to a linear program. For example, linear programming can be used to solve the least-absolute-values criterion or the minimax criterion. However, linear programming are considered to have no advantage over gradient-based methods (see Section 4.4.2 in [107]), and thus we do not discuss such methods here. Nevertheless, there are still many optimization algorithms that can be covered here, and we refer the readers to [13, 83].

For simplicity of presentation, we consider the problem of minimizing a function $g(\mathbf{m})$. We consider iterative algorithms which start with an iterate $\mathbf{m}_0$, and compute new iterates using

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \lambda_k p_k, \tag{16}$$

where p_k is a search direction, and λ_k a step size. Unless otherwise stated, we focus on unconstrained optimization. These algorithms can be used to directly solve the inverse problem in Eq. (5). We only present a selected subset of the algorithms available and have to omit many other interesting algorithms.

Newton-type methods. The classical Newton's method starts with an initial iterate $\mathbf{m}_0$, and computes new iterates using

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \left(\nabla^2 g(\mathbf{m}_k)\right)^{-1} \nabla g(\mathbf{m}_k), \tag{17}$$

that is, the search direction is $p_k = -\left(\nabla^2 g(\mathbf{m}_k)\right)^{-1} \nabla g(\mathbf{m}_k)$, and the step length is $\lambda_k = 1$. The basic Newton's method has quadratic local convergence rate at a small neighborhood of a local minimum. However, computing the search direction p_k can be very expensive, and thus many variants have been developed. In addition, in non-convex problems, classical Newton direction might not exist (if the Hessian matrix is not invertible) or it might not be an appropriate direction for descent (if the Hessian matrix is not positive definite).

For non-linear least squares problems, where the objective function $g(\mathbf{m})$ is a sum of squares of nonlinear functions, the Gauss-Newton (GN) method is often used [104]. Extensions to more general objective functions as in Eq. (13) with covariance matrix Σ and arbitrary regularization as in Eq. (14) is considered in [97]. Without loss of generality, assume $g(\mathbf{m}) = \sum_{i=1}^{s} (\mathbf{f}_i(\mathbf{m}) - \mathbf{d}_i)^2$. At iteration k, the GN search direction p_k is given by

$$\left(\sum_{i=1}^{s} J_i^\top J_i\right) p_k = -\nabla g, \tag{18}$$

where the sensitivity matrix J_i and the gradient ∇g are given by

$$J_i = \frac{\partial \mathbf{f}_i}{\partial \mathbf{m}}(\mathbf{m}_k), \quad i = 1, \ldots, s, \tag{19}$$

$$\nabla g = 2 \sum_{i=1}^{s} J_i^T (\mathbf{f}_i(\mathbf{m}_k) - \mathbf{d}_i), \tag{20}$$

The Gauss-Newton method can be seen as an approximation of the basic Newton's method obtained by replacing $\nabla^2 g$ by $\sum_{i=1}^{s} J_i^\top J_i$. The step length $\lambda_k \in [0,1]$ can be determined by a weak line search [83] (using, say, the Armijo algorithm starting with $\lambda_k = 1$) ensuring sufficient decrease in $g(\mathbf{m}_{k+1})$ as compared to $g(\mathbf{m}_k)$.

Often several nontrivial modifications are required to adapt this prototype method for different applications, e.g., dynamic regularization [31,56,89,90] and more general *stabilized GN* studied [33,94]. This method replaces the solution of the linear systems defining p_k by r preconditioned conjugate gradient (PCG) inner iterations, which costs $2r$ solutions of the forward problem per iteration, for a moderate integer value r. Thus, if K outer iterations are required to obtain an acceptable solution then the total work estimate (in terms of the number of PDE solves) is approximated *from below* by $2(r+1)Ks$.

Though Gauss-Newton is arguable the method of choice within the inverse problem community, other Newton-type methods exist which have been designed to suitably deal with the non-convex nature of the underlying optimization problem include Trust Region [27,113] and the Cubic Regularization [23,113]. These methods have recently found applications in machine learning [114]. Studying the advantages/disadvantages of these non-convex methods for solving inverse problems can be indeed a useful undertaking.

Quasi-Newton methods. An alternative method to the above Newton-type methods is the quasi-Newton variants including the celebrated limited memory BFGS (L-BFGS) [71,82]. BFGS iteration is closely related to conjugate gradient (CG) iteration. In particular, BFGS applied to a strongly convex quadratic objective, with exact line search as well as initial Hessian P, is equivalent to preconditioned CG with preconditioner P. However, as the objective function departs from being a simple quadratic, the number of iterations of L-BFGS could be significantly higher than that of GN or trust region. In addition, it has been shown that the performance of BFGS and its limited memory version is greatly negatively affected by the high degree if ill-conditioning present in such problems [95,96,115]. These two factor are among the main reasons why BFGS (and L-BFGS) can be less effective compared with other Newton-type alternatives in many inversion applications [47].

Krylov subspace method. A Krylov subspace method iteratively finds the optimal solution to an optimization in a larger subspace by making use of the previous solution in a smaller subspace. One of the most commonly used Krylov subspace method is the conjugate gradient (CG) method. CG was originally designed to solve convex quadratic minimization problems of the form $g(\mathbf{m}) = \frac{1}{2}\mathbf{m}^\top A\mathbf{m} - b^\top \mathbf{m}$. Equivalently, this solves the positive definite linear system $A\mathbf{m} = b$. It computes a sequence of iterates $\mathbf{m}_0, \mathbf{m}_1, \ldots$ converging to the minimum through the following two set of equations.

$$\mathbf{m}_0 = 0, \qquad r_0 = b, \qquad p_0 = r_0, \tag{21}$$

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \frac{||r_k||_2^2}{p_k^\top A p_k} p_k, \quad r_{k+1} = r_k - \frac{||r_k||_2^2}{p_k^\top A p_k} A p_k, \quad p_{k+1} = r_{k+1} + \frac{||r_{k+1}||_2^2}{||r_k||_2^2} p_k, \quad k \geq 0. \tag{22}$$

This can be used to solve the forward problem of the form $L_i(\mathbf{m})\mathbf{u}_i = \mathbf{q}_i$, provided that $L_i(\mathbf{m})$ is positive definite, which is true in many cases.

CG can be used to solve the linear system for the basic Newton direction. However, the Hessian is not necessarily positive definite and modification is needed [83].

In general, CG can be generalized to minimize a nonlinear function $g(\mathbf{m})$ [28,42]. It starts with an arbitrary $\mathbf{m}_0$, and $p_1 = -\nabla g(\mathbf{m}_0)$, and computes a sequence of iterates $\mathbf{m}_1, \mathbf{m}_2, \ldots$ using the equations below: for $k \geq 0$,

$$\mathbf{m}_{k+1} = \arg \min_{\mathbf{m} \in \{\mathbf{m}_k + \lambda p_k, \lambda \in \mathbb{R}\}} g(\mathbf{m}), \tag{23}$$

$$p_{k+1} = -\nabla g(\mathbf{m}_{k+1}) + \beta_k p_k, \qquad \text{where } \beta_k = \frac{||\nabla g(\mathbf{m}_{k+1})||_2^2}{||\nabla g(\mathbf{m}_k)||_2^2}. \tag{24}$$

The above formula for β_k is known as the Fletcher-Reeves formula. Other choices of β_k exist. The following two formula are known as the Polak-Ribiere and Hestenes-Stiefel formula respectively.

$$\beta_k = \frac{\langle \nabla g(\mathbf{m}_{k+1}) - \nabla g(\mathbf{m}_k), \nabla g(\mathbf{m}_{k+1}) \rangle}{||\nabla g(\mathbf{m}_k)||_2^2}, \tag{25}$$

$$\beta_k = \frac{\langle \nabla g(\mathbf{m}_{k+1}) - \nabla g(\mathbf{m}_k), \nabla g(\mathbf{m}_{k+1}) \rangle}{p_k^\top (\nabla g(\mathbf{m}_{k+1}) - \nabla g(\mathbf{m}_k))}. \tag{26}$$

In practice, nonlinear CG does not seem to work well, and is mainly used together with other methods, such as in the Newton CG method [83].

Lagrangian method of multipliers. The above discussion focuses on unconstrained optimization algorithms, which are suitable for unconstrained formulations of inverse problems, or unconstrained auxiliary optimization problems in methods which solves the constrained formulations directly. The Lagrangian method of multipliers is often used to directly solve the constrained version. Algorithms have been developed to offset the heavier computational cost and slow convergence rates of standard algorithms observed on the Lagrangian, which is a larger problem than the constrained problem. For example, such algorithm may reduce the problem to a smaller one, such as working with the reduced Hessian of the Lagrangian [50], or preconditioning [10, 49]. These methods have shown some success in certain PDE-constrained optimization problems.

Augmented Lagrangian methods have also been developed (e.g. [1, 60]). Such method constructs a series of penalized Lagrangians with vanishing penalty, and finds an optimizer of the Lagrangian by successively optimizing the penalized Lagrangians.

2.5 Challenges

Scaling up to large problems. The discretized version of an inverse problem is usually of very large scale, and working with fine resolution or discretized problems in high dimension is still an active area of research.

Another challenge is to scale up to large number of measurements, which is widely believed to be helpful for quality reconstruction of the model in practice, with some theoretical support. While recent technological advances makes many big datasets available, existing algorithms cannot efficiently cope with such datasets. Examples of such problems include electromagnetic data inversion in mining exploration [36, 48, 81, 84], seismic data inversion in oil exploration [41, 59, 91], diffuse optical tomography (DOT) [6, 14], quantitative photo-acoustic tomography (QPAT) [45, 117], direct current (DC) resistivity [33, 52, 54, 86, 103], and electrical impedance tomography (EIT) [16, 24, 110].

It has been suggested that many well-placed experiments yield practical advantage in order to obtain reconstructions of acceptable quality. For the special case where the measurement locations as well as the discretization matrices do not change from one experiment to another, various approximation techniques have been proposed to reduce the effective number of measurements, which in turn implies a smaller scale optimization problem, under the unifying category of "simultaneous sources inversion" [51, 66, 92, 94, 97]. Under certain circumstances, even if the P_i's are different across experiments (but L_i's are fixed), there are methods to transform the existing data set into the one where all sources share the same receivers, [93].

Dealing with non-convexity. Another major source of difficulty in solving many inverse problems, is the high-degree of non-linearity and non-convexity in (1). This is most often encountered in problems involving PDE-constrained optimization where each $\mathbf{f}_i$ corresponds to the solution of a PDE. Even if the output of the PDE model itself, i.e., the "right-hand side", is linear in the sought-after parameter, the solution of the PDE, i.e., the forward problem, shows a great deal of non-linearity. This coupled with a great amount of non-convexity can have significant consequences in the quality of inversion and the obtained parameter. Indeed, in presence of non-convexity, the large-scale computational challenges are exacerbated, multiple folds over, by the difficulty of avoiding (possibly degenerate) *saddle-points* as well as finding (at least) a *local minimum*.

Dealing with discontinuity. While the parameter function of the model is often smooth, the parameter function can be discontinuous in some cases. Such discontinuities arise very naturally as a result of the physical properties of the underlying physical system, e.g., EIT and DC resistivity, and require non-trivial modifications to optimization algorithms, e.g., [33, 94]. Ignoring such discontinuities can lead to unsatisfactory recovery results [33, 34, 106]. The level set method [85] is often used to model discontinuous parameter function. This reparametrizes the discontinuous parameter function as a differentiable one, and thus enabling more stable optimization [34].

3 Recent Advances in Optimization

Recent successes in using machine learning to deal with challenging perception and natural language understanding problems have spurred many advances in the study of optimization algorithms as optimization is a building block in machine learning. These new developments include efficient methods for large-scale optimization, methods designed to handle non-convex problems, methods incorporating the structural constraints, and finally the revival of second-order methods. While these developments address a different set of applications in machine learning, they address similar issues as encountered in inverse optimization and could be useful. We highlight some of the works below. We keep the discussion brief because numerous works have been done behind these developments and an indepth and comprehensive discussion is beyond the scope of this review. Our objective is thus to delineate the general trends and ideas, and provide references for interested readers to dig on relevant topics.

Stochastic optimization. The development in large-scale optimization methods is driven by the availability of many large datasets, which are made possible by the rapid development and extensive use of computers and information technology. In machine learning, a model is generally built by optimizing a sum of misfit on the examples. This finite-sum structure naturally invites the application of stochastic optimization algorithms. This is mainly due to the fact that stochastic algorithms recover the sought-after models more efficiently by employing small batches of data in each iteration, as opposed to the whole data-set. The most well-known stochastic gradient based algorithm is the stochastic gradient descent (SGD). To minimize a finite-sum objective function

$$g(\mathbf{m}) = \frac{1}{n}\sum_{i=1}^{n} g_i(\mathbf{m}), \tag{27}$$

in the big data regime where $n \gg 1$, the vanilla SGD performs an update

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \lambda_k \nabla g_{i_k}(\mathbf{m}_k), \tag{28}$$

where i_k is randomly sampled from $1,\ldots,n$. As compared to gradient descent, SGD replaces the full gradient $\nabla g(\mathbf{m})$ by a stochastic gradient $g_{i_k}(\mathbf{m}_k)$ with its expectation being the full gradient. The batch version of SGD constructs a stochastic gradient by taking the average of several stochastic gradients.

Vanilla SGD is inexpensive per iteration, but suffers from a slow rate of convergence. For example, while full gradient descent achieves a linear convergence rate for smooth strongly convex problems, SGD only converges at a sublinear rate. The slow convergence rate can be partly accounted by the variance in the stochastic gradient. Recently, variance reduction techniques have been developed, e.g. SVRG [64] and SDCA [102]. Perhaps surprisingly, such variants can achieve linear convergence rates on convex smooth problems as full gradient descent does, instead of sublinear rates achieved by the vanilla SGD. There are also a number of variants

with no known linear rates but have fast convergence rates for non-convex problems in practice, e.g., AdaGrad [37], RMSProp [108], ESGD [29], Adam [65], and Adadelta [118]. Indeed, besides efficiency, stochastic optimization algorithms also seem to be able to cope with the nonconvex objective functions well, and play a key role in the revival of neural networks as deep learning [46, 63, 69].

Recently, it has also been shown that SGD can be used as a variational algorithm for computing the posterior distribution of parameters given observations [75]. This can be useful in the Bayesian approach for solving inverse problems.

Nonconvex optimization. There is also an increasing interest in non-convex optimization in the machine learning community recently. Nonconvex objectives not only naturally occur in deep learning, but also occur in problems such as tensor decomposition, variable selection, low-rank matrix completion, e.g. see [46, 62, 76] and references therein.

As discussed above, stochastic algorithms have been found to be capable of effectively escaping local minima. There are also a number of studies which adapt well-known acceleration techniques for convex optimization to accelerate the convergence rates of both stochastic and non-stochastic optimization algorithms for nonconvex problems, e.g., [4, 70, 88, 105].

Dealing with structural constraints. Many problems in machine learning come with complex structural constraints. The Frank-Wolfe algorithm (a.k.a. conditional gradient) [44] is an algorithm for optimizing over a convex domain. It has gained a revived interest due to its ability to deal with many structural constraints efficiently. It requires solving a linear minimization problem over the feasible set, instead of a quadratic program as in the case of proximal gradient algorithms or projected gradient descent. Domains suitable for the Frank-Wolfe algorithm include simplices, ℓ_p-balls, matrix nuclear norm ball, matrix operator norm ball [61].

The Frank-Wolfe algorithm belongs to the class of linear-optimization-based algorithms [67, 68]. These algorithms share with the Frank-Wolfe algorithm the characteristic of requiring a first-order oracle for gradient computation and an oracle for solving a linear optimization problem over the constraint set.

Second-order methods. The great appeal of the second-order methods lies mainly in the observed empirical performance as well as some very appealing theoretical properties. For example, it has been shown that stochastic Newton-type methods in general, and Gauss-Newton in particular, can not only be made scalable and have low per-iteration cost [33, 50, 53, 93, 94, 97], but more importantly, and unlike first-order methods, are very *resilient* to many adversarial effects such as *ill-conditioning* [95, 96, 115]. As a result, for moderately to very ill-conditioned problems, commonly found in scientific computing, while first-order methods make effectively no progress at all, second-order counterparts are not affected by the degree of ill-conditioning. A more subtle, yet potentially more severe draw-back in using first-order methods, is that their success is tightly intertwined with *fine-tunning* (often many) *hyper-parameters*, most importantly, the step-size [11]. In fact, it is highly unlikely that many of these methods exhibit acceptable performance on first try, and

it often takes many trials and errors before one can see reasonable results. In contrast, second-order optimization algorithms involve much less parameter tuning and are less sensitive to the choice of hyper-parameters [11, 114].

Since for the finite-sum problem (27) with $n \gg 1$, the operations with the Hessian/gradient constitute major computational bottlenecks, a rather more recent line of research is to construct the inexact Hessian information using the application of *randomized methods*. Specifically, for convex optimization, the stochastic approximation of the full Hessian matrix in the classical Newton's method has been recently considered in [3, 11, 15, 19, 20, 38, 40, 78, 79, 87, 95, 96, 112, 115, 116]. In addition to inexact Hessian, a few of these methods study the fully stochastic case in which the gradient is also approximated, e.g., [15, 95, 96]. For non-convex problems, however, the literature on methods that employ randomized Hessian approximation is significantly less developed than that of convex problems. A few recent examples include the stochastic trust region [113], stochastic cubic regularization [109, 113], and noisy negative curvature method [72]. Empirical performance of many of these methods for some non-convex machine learning applications has been considered in [114].

3.1 A concrete success story

The development of optimization methods in the machine learning community has been fueled by the need to obtain better generalization performance one future "unseen" data. This is in contrast with typical inverse problem applications where fitting the model to the observations on hand make up of all that matters. These rather strikingly different goals have lead the ML community to develop optimization methods that can address ML specific challenges. This, in part, has given rise to scalable algorithms that can often deliver far beyond what the most widely used optimization methods in the inverse problem community can.

As a concrete example, consider L-BFGS and Gauss-Newton, which are, arguably, among the most popular optimization techniques used by the scientific computing community in a variety of inverse problem applications. In fact, unlike Gauss-Newton method, L-BFGS, due to its low per-iteration costs, has found significant attraction within the machine learning community as well. Nevertheless, due to the resurgence of non-convex deep learning problems in ML, there is an increasing demand for scalable optimization algorithms that can avoid saddle points and converge to a local minimum. This demand has driven the development algorithms that can surpass the performance of L-BFGS and Gauss-Newton when applied to deep learning applications, e.g., [114].

These results are not unexpected. Indeed, contrary to popular belief, BFGS is not quite a "full-fledged" second-order method as it merely employs first-order information, i.e. gradients, to approximate the curvature. Similar in spirit, Gauss-Newton also does not fully utilize the Hessian information. In particular, in exchange for obtaining a positive definite approximation matrix, GN completely ignores the in-

formation from *negative curvature*, which is critical for allowing to escape from regions with small gradient. Escaping saddle points and converging to a local minimum with lower objective values have surprisingly not been a huge concern for the inverse problem community. This is in sharp contrast to the machine learning applications where obtaining lower training errors with deep learning models typically translates to better generalization performance.

4 Discussion

Optimization is not only used in the data fitting approach to inverse problems, but also used in the Bayesian approach. An important problem in the Bayesian approach is the choice of the parameters for the prior. While these were often chosen in a somewhat ad hoc way, there are studies which use sampling [2, 43], hierarchical prior models [21, 22], and optimization [9, 73] methods to choose the parameters. While choosing the prior parameters through optimization has found some success, such optimization is hard and it remains a challenge to develop effective algorithms to solve these problems.

For inverse problems with large number of measurements, solving each forward problem can be expensive, and the mere evaluation of the misfit function may become computationally prohibitive. Stochastic optimization algorithms might be beneficial in this case, because the objective function is often a sum of misfits over different measurements.

The data fitting problem is generally non-convex and thus optimization algorithms may be trapped in a local optimum. Stochastic optimization algorithms also provide a means to escape the local optima. Recent results in nonconvex optimization, such as those on accelerated methods, may provide more efficient alternatives to solve the data fitting problem.

While box constraints are often used in inverse problems because they are easier to deal with, simplex constraint can be beneficial. The Frank-Wolfe algorithm provides a efficient way to deal with the simplex constraint, and can be a useful tool to add on to the toolbox of an inverse problem researcher.

5 Conclusion

State-of-the-art optimization methods in the inverse problem community struggle to cope with important issues such as large-scale problems and nonconvexity. At the same time, many progresses in optimization have been made in the machine learning community. Our discussion on the connections has been brief. Nevertheless, we have highlighted the valuable potential synergies that are to be reaped by bringing these two communities closer together.

Acknowledgement. We thank the anonymous reviewers for their helpful comments.

References

1. Abdoulaev, G.S., Ren, K., Hielscher, A.H.: Optical tomography as a PDE-constrained optimization problem. Inverse Problems **21**(5), 1507 – 1530 (2005)
2. Agapiou, S., Bardsley, J.M., Papaspiliopoulos, O., Stuart, A.M.: Analysis of the gibbs sampler for hierarchical inverse problems. SIAM/ASA Journal on Uncertainty Quantification **2**(1), 511–544 (2014)
3. Agarwal, N., Bullins, B., Hazan, E.: Second order stochastic optimization in linear time. arXiv preprint arXiv:1602.03943 (2016)
4. Allen-Zhu, Z., Hazan, E.: Variance reduction for faster non-convex optimization. arXiv preprint arXiv:1603.05643 (2016)
5. Archer, G., Titterington, D.: On some bayesian/regularization methods for image restoration. Image Processing, IEEE Transactions on **4**(7), 989–995 (1995)
6. Arridge, S.R.: Optical tomography in medical imaging. Inverse problems **15**(2), R41 (1999)
7. Arridge, S.R., Hebden, J.C.: Optical imaging in medicine: Ii. modelling and reconstruction. Physics in Medicine and Biology **42**(5), 841 (1997)
8. Aster, R.C., Borchers, B., Thurber, C.H.: Parameter estimation and inverse problems. Academic Press (2013)
9. Bardsley, J.M., Calvetti, D., Somersalo, E.: Hierarchical regularization for edge-preserving reconstruction of pet images. Inverse Problems **26**(3), 035,010 (2010)
10. Benzi, M., Haber, E., Taralli, L.: A preconditioning technique for a class of pde-constrained optimization problems. Advances in Computational Mathematics **35**(2), 149–173 (2011)
11. Berahas, A.S., Bollapragada, R., Nocedal, J.: An Investigation of Newton-Sketch and Subsampled Newton Methods. arXiv preprint arXiv:1705.06211 (2017)
12. Bertero, M., Boccacci, P.: Introduction to inverse problems in imaging. CRC press (2010)
13. Björck, Å.: Numerical methods for least squares problems. SIAM (1996)
14. Boas, D., Brooks, D., Miller, E., DiMarzio, C.A., Kilmer, M., Gaudette, R., Zhang, Q.: Imaging the body with diffuse optical tomography. Signal Processing Magazine, IEEE **18**(6), 57–75 (2001)
15. Bollapragada, R., Byrd, R., Nocedal, J.: Exact and inexact subsampled Newton methods for optimization. arXiv preprint arXiv:1609.08502 (2016)
16. Borcea, L., Berryman, J.G., Papanicolaou, G.C.: High-contrast impedance tomography. Inverse Problems **12**, 835–858 (1996)
17. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. arXiv preprint arXiv:1606.04838 (2016)
18. Bunks, C., Saleck, F.M., Zaleski, S., Chavent, G.: Multiscale seismic waveform inversion. Geophysics **60**(5), 1457–1473 (1995)
19. Byrd, R.H., Chin, G.M., Neveitt, W., Nocedal, J.: On the use of stochastic Hessian information in optimization methods for machine learning. SIAM Journal on Optimization **21**(3), 977–995 (2011)
20. Byrd, R.H., Chin, G.M., Nocedal, J., Wu, Y.: Sample size selection in optimization methods for machine learning. Mathematical programming **134**(1), 127–155 (2012)
21. Calvetti, D., Somersalo, E.: A gaussian hypermodel to recover blocky objects. Inverse problems **23**(2), 733 (2007)
22. Calvetti, D., Somersalo, E.: Hypermodels in the bayesian imaging framework. Inverse Problems **24**(3), 034,013 (2008)
23. Cartis, C., Gould, N.I., Toint, P.L.: Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization. Optimization Methods and Software **27**(2), 197–219 (2012)

24. Cheney, M., Isaacson, D., Newell, J.C.: Electrical impedance tomography. SIAM Review **41**, 85–101 (1999)
25. Claerbout, J.F., Muir, F.: Robust modeling with erratic data. Geophysics **38**(5), 826–844 (1973)
26. Clason, C.: L∞ fitting for inverse problems with uniform noise. Inverse Problems **28**(10), 104,007 (2012)
27. Conn, A.R., Gould, N.I., Toint, P.L.: Trust region methods, vol. 1. SIAM (2000)
28. Dai, Y.: Nonlinear conjugate gradient methods. Wiley Encyclopedia of Operations Research and Management Science (2011)
29. Dauphin, Y., de Vries, H., Bengio, Y.: Equilibrated adaptive learning rates for non-convex optimization. In: Advances in Neural Information Processing Systems, pp. 1504–1512 (2015)
30. van den Doel, K., Ascher, U.M.: On level set regularization for highly ill-posed distributed parameter estimation problems. J. Comp. Phys. **216**, 707–723 (2006)
31. van den Doel, K., Ascher, U.M.: Dynamic level set regularization for large distributed parameter estimation problems. Inverse Problems **23**, 1271–1288 (2007)
32. van den Doel, K., Ascher, U.M.: Dynamic regularization, level set shape optimization, and computed myography. Control and Optimization with Differential-Algebraic Constraints **23**, 315 (2012)
33. Doel, K.v.d., Ascher, U.: Adaptive and stochastic algorithms for EIT and DC resistivity problems with piecewise constant solutions and many measurements. SIAM J. Scient. Comput. **34**, DOI: 10.1137/110826,692 (2012)
34. Doel, K.v.d., Ascher, U., Leitao, A.: Multiple level sets for piecewise constant surface reconstruction in highly ill-posed problems. Journal of Scientific Computation **43**(1), 44–66 (2010)
35. Domingos, P.: A few useful things to know about machine learning. Communications of the ACM **55**(10), 78–87 (2012)
36. Dorn, O., Miller, E.L., Rappaport, C.M.: A shape reconstruction method for electromagnetic tomography using adjoint fields and level sets. Inverse Problems **16** (2000). 1119-1156
37. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. The Journal of Machine Learning Research **12**, 2121–2159 (2011)
38. Eisen, M., Mokhtari, A., Ribeiro, A.: Large Scale Empirical Risk Minimization via Truncated Adaptive Newton Method. arXiv preprint arXiv:1705.07957 (2017)
39. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Dordrecht (1996)
40. Erdogdu, M.A., Montanari, A.: Convergence rates of sub-sampled newton methods. In: Advances in Neural Information Processing Systems 28, pp. 3034–3042 (2015)
41. Fichtner, A.: Full Seismic Waveform Modeling and Inversion. Springer (2011)
42. Fletcher, R.: Practical methods of optimization. John Wiley & Sons (2013)
43. Fox, C., Norton, R.A.: Fast sampling in a linear-gaussian inverse problem. SIAM/ASA Journal on Uncertainty Quantification **4**(1), 1191–1218 (2016)
44. Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval research logistics quarterly **3**(1-2), 95–110 (1956)
45. Gao, H., Osher, S., Zhao, H.: Quantitative photoacoustic tomography. In: Mathematical Modeling in Biomedical Imaging II, pp. 131–158. Springer (2012)
46. Ge, R., Huang, F., Jin, C., Yuan, Y.: Escaping from saddle points-online stochastic gradient for tensor decomposition. In: COLT, pp. 797–842 (2015)
47. Haber, E.: Quasi-newton methods for large-scale electromagnetic inverse problems. Inverse problems **21**(1), 305 (2004)
48. Haber, E., Ascher, U., Oldenburg, D.: Inversion of 3D electromagnetic data in frequency and time domain using an inexact all-at-once approach. Geophysics **69**, 1216–1228 (2004)
49. Haber, E., Ascher, U.M.: Preconditioned all-at-once methods for large, sparse parameter estimation problems. Inverse Problems **17**(6), 1847 (2001)
50. Haber, E., Ascher, U.M., Oldenburg, D.: On optimization techniques for solving nonlinear inverse problems. Inverse problems **16**(5), 1263 (2000)

51. Haber, E., Chung, M.: Simultaneous source for non-uniform data variance and missing data. arXiv preprint arXiv:1404.5254 (2014)
52. Haber, E., Chung, M., Herrmann, F.: An effective method for parameter estimation with PDE constraints with multiple right-hand sides. SIAM J. Optimization **22**, 739–757 (2012)
53. Haber, E., Chung, M., Herrmann, F.: An effective method for parameter estimation with PDE constraints with multiple right-hand sides. SIAM Journal on Optimization **22**(3), 739–757 (2012)
54. Haber, E., Heldmann, S., Ascher, U.: Adaptive finite volume method for distributed non-smooth parameter identification. Inverse Problems **23**, 1659–1676 (2007)
55. Hadamard, J.: Sur les problèmes aux dérivées partielles et leur signification physique. Princeton University Bulletin pp. 49 – 52 (1902)
56. Hanke, M.: Regularizing properties of a truncated Newton-cg algorithm for nonlinear inverse problems. Numer. Funct. Anal. Optim. **18**, 971–993 (1997)
57. Hansen, P.C.: Rank-Deficient and Discrete Ill-Posed Problems. SIAM (1998)
58. Herman, G.T.: Fundamentals of computerized tomography: image reconstruction from projections. Springer Science & Business Media (2009)
59. Herrmann, F., Erlangga, Y., Lin, T.: Compressive simultaneous full-waveform simulation. Geophysics **74**, A35 (2009)
60. Ito, K., Kunisch, K.: The augmented lagrangian method for parameter estimation in elliptic systems. SIAM Journal on Control and Optimization **28**(1), 113–136 (1990)
61. Jaggi, M.: Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13), pp. 427–435 (2013)
62. Jain, P., Netrapalli, P., Sanghavi, S.: Low-rank matrix completion using alternating minimization. In: Proceedings of the forty-fifth annual ACM symposium on Theory of computing, pp. 665–674. ACM (2013)
63. Jin, C., Ge, R., Netrapalli, P., Kakade, S.M., Jordan, M.I.: How to escape saddle points efficiently. arXiv preprint arXiv:1703.00887 (2017)
64. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in Neural Information Processing Systems, pp. 315–323 (2013)
65. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
66. Kumar, R., Silva, C.D., Akalin, O., Aravkin, A.Y., Mansour, H., Recht, B., Herrmann, F.J.: Efficient matrix completion for seismic data reconstruction (2014). Submitted to Geophysics on August 8, 2014.
67. Lan, G., Pokutta, S., Zhou, Y., Zink, D.: Conditional accelerated lazy stochastic gradient descent. In: ICML. PMLR (2017). URL http://proceedings.mlr.press/v70/lan17a.html
68. Lan, G., Zhou, Y.: Conditional gradient sliding for convex optimization. SIAM Journal on Optimization **26**(2), 1379–1409 (2016)
69. Levy, K.Y.: The Power of Normalization: Faster Evasion of Saddle Points. arXiv preprint arXiv:1611.04831 (2016)
70. Li, H., Lin, Z.: Accelerated proximal gradient methods for nonconvex programming. In: Advances in neural information processing systems, pp. 379–387 (2015)
71. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Mathematical programming **45**(1-3), 503–528 (1989)
72. Liu, M., Yang, T.: On Noisy Negative Curvature Descent: Competing with Gradient Descent for Faster Non-convex Optimization. arXiv preprint arXiv:1709.08571 (2017)
73. Liu, W., Li, J., Marzouk, Y.M.: An approximate empirical bayesian method for large-scale linear-gaussian inverse problems. arXiv preprint arXiv:1705.07646 (2017)
74. Louis, A.: Medical imaging: state of the art and future development. Inverse Problems **8**(5), 709 (1992)
75. Mandt, S., Hoffman, M., Blei, D.: A variational analysis of stochastic gradient algorithms. In: International Conference on Machine Learning, pp. 354–363 (2016)

76. Mazumder, R., Friedman, J.H., Hastie, T.: Sparsenet: Coordinate descent with nonconvex penalties. Journal of the American Statistical Association **106**(495), 1125–1138 (2011)
77. Menke, W.: Geophysical data analysis: discrete inverse theory. Academic press (2012)
78. Mutnỳ, M.: Stochastic Second-Order Optimization via von Neumann Series. arXiv preprint arXiv:1612.04694 (2016)
79. Mutnỳ, M., Richtárik, P.: Parallel Stochastic Newton Method. arXiv preprint arXiv:1705.02005 (2017)
80. Natterer, F., Wübbeling, F.: Mathematical methods in image reconstruction. SIAM (2001)
81. Newman, G.A., Alumbaugh, D.L.: Frequency-domain modelling of airborne electromagnetic responses using staggered finite differences. Geophys. Prospecting **43**, 1021–1042 (1995)
82. Nocedal, J.: Updating quasi-Newton matrices with limited storage. Mathematics of computation **35**(151), 773–782 (1980)
83. Nocedal, J., Wright, S.: Numerical optimization. Springer Science & Business Media (2006)
84. Oldenburg, D., Haber, E., Shekhtman, R.: 3D inverseion of multi-source time domain electromagnetic data. J. Geophysics (2013). To appear
85. Osher, S., Sethian, J.: Fronts propagating with curvature dependent speed: algorithms based on Hamilton-Jacobi formulations. J. Comp. Phys. **79**, 12–49 (1988)
86. Pidlisecky, A., Haber, E., Knight, R.: RESINVM3D: A MATLAB 3D Resistivity Inversion Package. Geophysics **72(2)**, H1–H10 (2007)
87. Pilanci, M., Wainwright, M.J.: Newton sketch: A linear-time optimization algorithm with linear-quadratic convergence. arXiv preprint arXiv:1505.02250 (2015)
88. Reddi, S.J., Hefny, A., Sra, S., Poczos, B., Smola, A.: Stochastic variance reduction for nonconvex optimization. arXiv preprint arXiv:1603.06160 (2016)
89. Rieder, A.: Inexact Newton regularization using conjugate gradients as inner iteration. SIAM J. Numer. Anal. **43**, 604–622 (2005)
90. Rieder, A., Lechleiter, A.: Towards a general convergence theory for inexact Newton regularizations. Numer. Math. **114**(3), 521–548 (2010)
91. Rohmberg, J., Neelamani, R., Krohn, C., Krebs, J., Deffenbaugh, M., Anderson, J.: Efficient seismic forward modeling and acquisition using simultaneous random sources and sparsity. Geophysics **75**(6), WB15–WB27 (2010)
92. Roosta-Khorasani, F.: Randomized algorithms for solving large scale nonlinear least squares problems. Ph.D. thesis, University of British Columbia (2015)
93. Roosta-Khorasani, F., van den Doel, K., Ascher, U.: Data completion and stochastic algorithms for PDE inversion problems with many measurements. Electronic Transactions on Numerical Analysis **42**, 177–196 (2014)
94. Roosta-Khorasani, F., van den Doel, K., Ascher, U.: Stochastic algorithms for inverse problems involving PDEs and many measurements. SIAM J. Scientific Computing **36**(5), S3–S22 (2014). DOI 10.1137/130922756
95. Roosta-Khorasani, F., Mahoney, M.W.: Sub-sampled Newton methods I: Globally convergent algorithms. arXiv preprint arXiv:1601.04737 (2016)
96. Roosta-Khorasani, F., Mahoney, M.W.: Sub-sampled Newton methods II: Local convergence rates. arXiv preprint arXiv:1601.04738 (2016)
97. Roosta-Khorasani, F., Székely, G.J., Ascher, U.: Assessing stochastic algorithms for large scale nonlinear least squares problems using extremal probabilities of linear combinations of gamma random variables. SIAM/ASA Journal on Uncertainty Quantification **3**(1), 61–90 (2015)
98. w. Rundell, Engl, H.W.: Inverse problems in medical imaging and nondestructive testing. Springer-Verlag New York, Inc. (1997)
99. Russell, B.H.: Introduction to seismic inversion methods, vol. 2. Society of Exploration Geophysicists (1988)
100. Scharf, L.L.: Statistical signal processing, vol. 98. Addison-Wesley Reading, MA (1991)
101. Sen, M.K., Stoffa, P.L.: Global optimization methods in geophysical inversion. Cambridge University Press (2013)
102. Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss. The Journal of Machine Learning Research **14**(1), 567–599 (2013)

103. Smith, N.C., Vozoff, K.: Two dimensional DC resistivity inversion for dipole dipole data. IEEE Trans. on geoscience and remote sensing **GE 22**, 21–28 (1984)
104. Sun, W., Yuan, Y.X.: Optimization theory and methods: nonlinear programming, vol. 1. Springer Science & Business Media (2006)
105. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: International conference on machine learning, pp. 1139–1147 (2013)
106. Tai, X.C., Li, H.: A piecewise constant level set method for elliptic inverse problems. Appl. Numer. Math. **57**, 686–696 (2007)
107. Tarantola, A.: Inverse problem theory and methods for model parameter estimation. SIAM (2005)
108. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning **4** (2012)
109. Tripuraneni, N., Stern, M., Jin, C., Regier, J., Jordan, M.I.: Stochastic Cubic Regularization for Fast Nonconvex Optimization. arXiv preprint arXiv:1711.02838 (2017)
110. Van Den Doel, K., Ascher, U., Haber, E.: The lost honour of ℓ_2-based regularization. Radon Series in Computational and Applied Math (2013)
111. Vogel, C.: Computational methods for inverse problem. SIAM, Philadelphia (2002)
112. Wang, C.C., Huang, C.H., Lin, C.J.: Subsampled Hessian Newton methods for supervised learning. Neural computation (2015)
113. Xu, P., Roosta-Khorasani, F., Mahoney, M.W.: Newton-type methods for non-convex optimization under inexact hessian information. arXiv preprint arXiv:1708.07164 (2017)
114. Xu, P., Roosta-Khorasani, F., Mahoney, M.W.: Second-order optimization for non-convex machine learning: An empirical study. arXiv preprint arXiv:1708.07827 (2017)
115. Xu, P., Yang, J., Roosta-Khorasani, F., Ré, C., Mahoney, M.W.: Sub-Sampled Newton Methods with Non-Uniform Sampling. In: Advances In Neural Information Processing Systems (NIPS), pp. 2530–2538 (2016)
116. Ye, H., Luo, L., Zhang, Z.: Revisiting sub-sampled newton methods. arXiv preprint arXiv:1608.02875 (2016)
117. Yuan, Z., Jiang, H.: Quantitative photoacoustic tomography: Recovery of optical absorption coefficient maps of heterogeneous media. Applied physics letters **88**(23), 231,101–231,101 (2006)
118. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)