AVERAGING RANDOM PROJECTION: A FAST ONLINE SOLUTION FOR LARGE-SCALE CONSTRAINED STOCHASTIC OPTIMIZATION

Jialin Liu, Yuantao Gu, and Mengdi Wang†*

*Tsinghua National Laboratory for Information Science and Technology
Tsinghua University, Beijing 100084, CHINA
†Department of Operations Research and Financial Engineering
Princeton University, Princeton 08544, USA

ABSTRACT

Stochastic optimization finds wide application in signal processing, online learning, and network problems, especially problems processing large-scale data. We propose an *Incremental Constraint Averaging Projection Method* (ICAPM) that is tailored to optimization problems involving a large number of constraints. The ICAPM makes fast updates by taking sample gradients and averaging over random constraint projections. We provide a theoretical convergence and rate of convergence analysis for ICAPM. Our results suggests that averaging random projections significantly improves the stability of the solutions. For numerical tests, we apply the ICAPM to an online classification problem and a network consensus problem.

Index Terms— Stochastic Optimization, Large Scale Optimization, Random Projection Method, Incremental Constraint Projection Method

1. INTRODUCTION

Stochastic Optimization (SO) method is widely used in machine learning (see [3,4]), online learning (see [2,15,16]) and signal reconstruction (see [17–19]). It is designed to deal with objective function that involves stochastic noise or can be written as the sum of many component functions. In practical big-data-based problems, stochastic optimization method often exhibits fast or even optimal rate of convergence. However, most development on SO focused merely on unconstrained optimization problems.

In real world, most optimization problems are *constrained*, in which the decision variable x must lie within in a feasible set X. Feasibility problem is by itself an important subject in signal processing (see [20–22]). When the feasible set $X = \cap_{i=1}^m X_i$ involves many constraint sets, the optimization problems become substantially harder than unconstrained ones.

Motivated by these practical challenges with constraints, we consider the following problem (see [8, 9]):

$$\begin{aligned} &\min_x \ \Big\{F(x) = \mathbf{E}[f(x;v)]\Big\} \\ &\text{s.t. } x \in X = \cap_{i=1}^m X_i \end{aligned} \tag{1}$$

where $F : \Re^n \mapsto \Re$, $f(\cdot; v) : \Re^n \mapsto \Re$ are real-valued convex functions, the constraints X_i are convex closed sets, and v is a random variable. For solution of problem (1), one may consider the *Gradient Projection Method* (see [5–7]), in which each iteration makes a stochastic gradient update and takes a projection on X. This method has nice theoretical convergence guarantee. However, it runs into difficulty when X is as complicated as $X = \cap_{i=1}^m X_i$, in which case calculating the projection on X becomes computationally expensive.

For faster solution of problem (1) that accounts for the difficult constraints, an efficient solution is the *Incremental Constraint Projection Method* (ICPM) (see [8,9] and Algorithm 1). In its projection step, ICPM randomly picks an X_i from all constraints and takes a projection onto X_i. This allows the algorithm to process one constraint X_i at a time using cheap iteration, making it suitable for online learning. However, random constraint projection induces additional variance, which may adversely affect the algorithm's stability.

In this paper, we aim to improve the rate of convergence and stability of ICPM. We propose a new algorithm, namely the *Incremental Constraint Averaging Projection Method* (ICAPM). At each iteration, ICAMP processes a number of randomly selected constraints, and averages over multiple random constraint projection steps. We prove that the ICAMP converges almost surely to a global optimal solution, regardless of the initial solution. More importantly, we analyze the convergence rate and stability of the ICAMP. We show that the variance of solution is indeed reduced by averaged projection, and we show that the optimization error decreases at a rate of $O(1/k)$ with high probability. Then we apply the proposed ICAMP to an SVM problem and a network estimation problem, in which the algorithm's empirical performances strongly support our earlier analysis.

Outline Section 2 introduces the basic notation and reviews the ICPM algorithm. Section 3 introduces the ICAPM algorithm. Section 4 presents our main convergence and stability results for ICAPM. Section 5 describes two applications of ICAPM and section 6 gives the simulation results.

2. PRELIMINARIES

For solution of problem (1), we consider the simulation setting where:

- For a given x, we can obtain an unbiased sample gradient or subgradient of $F(x) = \mathbf{E}[f(x;v)]$.
- From the collection of constraints $X_1, \ldots, X_m$, we can obtain one or multiple constraint sets, which are sampled according to some probability distribution.

Under this simulation setting, the ICPM is given by Algorithm 1. We summarize the parameters and notations of Algorithm 1 as follows:

This work was supported by National 973 Program of China (Grant No. 2013CB329203) and National Natural Science Foundation of China (NSFC 61371137). The corresponding author of this paper is Yuantao Gu (gyt@tsinghua.edu.cn).

Algorithm 1: Incremental Constraint Projection Method (ICPM)

Choose an arbitary initial point $x_0 \in \Re^n$;
for $k = 0, 1, 2, \ldots$ **do**
(1) Sample a random (sub)gradient $\tilde{\nabla} f(x_k; v_k)$;
(2) Sample ω_k from the set of constraints $X_1, \ldots, X_m$.**(3)** Calculate x_{k+1} as:

$$x_{k+1} = \Pi_{\omega_k}[x_k - \alpha_k \widetilde{\nabla} f(x_k; v_k)] \; ; \tag{2}$$

- α_k denotes the step size. We often take $\alpha_k = k^{-\alpha}$, where $\alpha \in [0.5, 1]$.
- $\widetilde{\nabla} f(x; v) \in \partial_x f(x; v)$ denotes a subgradient of function $f(\cdot; v)$ at the point x.
- $\Pi_{\omega_k} y := \arg\min_{x \in \omega_k} \|x - y\|$ denotes the Euclidean projection of y on the convex set ω_k.

At every iteration, the ICPM processes a single sample gradient and a single sample constraint. Assuming that projecting onto a single X_i is easy, the ICPM can be implemented efficiently. For example, when X is a polyhedron and each X_i is a halfspace, the projection on X_i has a simple analytical form.

3. ICAPM ALGORITHM

In order to improve the stability of random projection, we propose the new algorithm ICAPM as given by Algorithm 2. At each iteration of ICAPM, we first take a stochastic gradient descent step starting from x_k, then we sample a number of constraints $\omega_{k,i}, i = 1, 2, \ldots, M_k$, and take the average of the projections as the next iterate x_{k+1}. It is easy to see that the proposed ICAPM contains the ICPM as a special case when $M_k = 1$ for all k. See Figure 1 for graphical visualization of the ICAPM procedure. By averaging over random projections, we expect to reduce the variance in iterates and to potentially improve the algorithm's convergence rate.

Intuitively, by taking average of random projections, we may reduce the variance at every iteration and keep the next iterate concentrated around its expectation (see Theorem 2 for a formal statement). As illustrated by Figure 1, this prevents the next iterate x_{k+1} from randomly jumping into a distant constraint set. While improving the stability of iterates, the averaging scheme is computationally efficient, as calculating each random projection still involves only one simple set X_i.

4. CONVERGENCE ANALYSIS

4.1. Convergence

Suppose there exists at least one optimal solution x^* to problem (1) to problem (1), i.e.,

$$\mathbf{E}[f(x; v)] \geq \mathbf{E}[f(x^*; v)], \;\; \forall\, x \in \cap_{i=1}^m X_i.$$

In this section, we consider the convergence of iterates x_k generated by ICAPM to the set of optimal solutions. We define

$$\mathscr{F}_k := \{v_t, \omega_{t,i}, x_t, y_t \mid |t = 1, 2, \ldots, k-1, i = 1, 2, \ldots, M_k\}$$

as the collection of random variables that are revealed up to the kth iterations. Before starting the analysis, we gives some basic assumptions on $F(x)$, $\{X_i\}$ and the sampling scheme.

Algorithm 2: Incremental Constraint Averaging Projection Method (ICAPM)

Choose an arbitary $x_0 \in \Re^n$ and positive integers $\{M_k\}$;
for $k = 0, 1, 2, \ldots$ **do**
(1) Sample a random (sub)gradient $\tilde{\nabla} f(x_k; v_k)$;
(2) Update using a gradient descent:

$$y_{k+1} = x_k - \alpha_k \widetilde{\nabla} f(x_k; v_k) \; ; \tag{3}$$

(3) Sample M_k constraints $\{\omega_{k,i}\}_{i=1}^{M_k}$ independently from $\{X_i\}_{i=1}^m$ according to a uniform distribution.
(4) Calculate x_{k+1} as the averages of random projections:

$$x_{k+1} = \frac{1}{M_k} \sum_{i=1}^{M_k} \Pi_{\omega_{k,i}} y_{k+1} \; ; \tag{4}$$

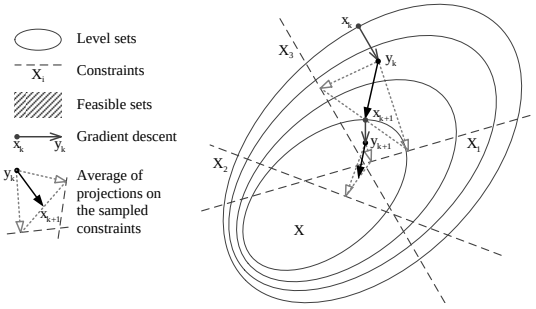

Fig. 1. Graphical visualization of the ICAPM algorithm

Assumption 1. *(Basic Assumptions)*

(a) The objective function $F(x)$ is convex and has bounded (sub)gradients.

(b) There exists a constant scalar η such that for any $x \in \Re^n$:

$$\|x - \Pi_X x\|^2 \leq \eta \max_{i=1,\ldots,m} \|x - \Pi_i x\|^2. \tag{5}$$

(c) The sample (sub)gradients are conditionally unbiased: for any $x \in \Re^n$,

$$\mathbf{E}[\tilde{\nabla} f(x, v_k) | \mathscr{F}_k] = \tilde{\nabla} F(x). \tag{6}$$

Our first main result establishes that the stochastic algorithm, ICAMP, converges with probability 1 to an global optimum, starting from an arbitrary initial solution.

Theorem 1 (Almost Sure Convergence)**.** *Suppose that the sequence $\{x_k\}$ is generated by ICAPM (Algorithm 2). Let Assumption 1 hold, and let the stepsize α_k satisfy:*

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \;\; \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{7}$$

Then $\{x_k\}$ converges almost surely to a random point in the set of optimal solutions for problem (1).

Proof. The key step is to analyze the relation between $\mathbf{E}[\|x_{k+1} - x^*\|^2|\mathscr{F}_k]$ and $\|x_k - x^*\|^2$. Then we use a *coupled supermartingale convergence theorem* (see Section 3 of [8]) to prove that x_k converges almost surely to some optimal solution. We refer to [1] for the complete proof analysis. □

4.2. Convergence Stability

We consider the stability of ICAMP in terms of the conditional variance $Var[x_{k+1}|\mathscr{F}_k]$ associated with each iteration. We obtain the following result:

Theorem 2 (Reduced Variance)**.** *Let Assumption 1 hold. Suppose that $\{x_k\}$ is generated by ICAPM (Algorithm 2), we can give an upper bound on the conditional variance per iteration:*

$$Var[x_{k+1}|\mathscr{F}_k] \leq \frac{1}{M_k} 16\|x_k - \Pi_X x_k\|^2 + \left(\frac{1}{M_k}64 + 2\right) D^2\alpha_k^2, \tag{8}$$

where D is a scalar such that $f(x_k, v_k) \leq D$ with probability 1, $\|x_k - \Pi_X x_k\|^2$ is the distance from x_k to the feasible set X, and M_k is the sample number of constraints at the kth iteration.

Proof. The key idea is to leverage the independency of multiple constraint sets $\omega_{k,i}$ to obtain a reduced variance. This analysis works for both ICPM and ICAPM. Please see [1] for the complete proof. □

This result implies that: the conditional variance diminishes to 0 as x_k converges, and more importantly, it is controllable via adjusting M_k. Note that this proposition also holds for ICPM when we let $M_k = 1$ for all k.

From Theorem 2, we notice that the conditional variance consists of two parts: the first part $(1/M_k)(16\|x_k - \Pi_X x_k\|^2 + 64D^2\alpha_k^2)$ is caused by randomness in selection of constraints ω_k, and the second part $2D^2\alpha_k^2$ is caused by randomness in the sample gradients $f(\cdot; v_k)$. By averaging over random projection, ICAPM is able to reduce the variance introduced by random constraints selection.

4.3. Convergence Rate with High Probability

Now we present our rate of convergence result. Assuming strong convexity and zero noise in gradients, we obtain the following error bound that holds with high probability.

Theorem 3 (Convergence Rate)**.** *Let Assumption 1 hold, let F be a σ-strongly convex function, and let $f(x; v) = F(x)$ for all v. Suppose that $\{x_k\}$ is generated by ICAPM with the stepsize $\alpha_k = \frac{1}{k+1}$, then there exist constants $C_1, C_2 > 0$ such that for any $T \geq 0$:*

$$\|x_k - x^*\|^2 \leq \frac{C_1}{k+1}, \quad \forall\, 0 \leq k \leq T, \tag{9}$$

with probability at least $\prod_{k=1}^{T}(1 - \frac{C_2}{M_k})$.

Proof. First we analyze the relation between $\mathbf{E}[\|x_{k+1} - x^*\|^2|\mathscr{F}_k]$ and $\|x_k - x^*\|^2$. Then we use the *Chebyshelv Inequality* to obtain an inequality relation between $\|x_{k+1} - x^*\|^2$ and $\|x_k - x^*\|^2$ that holds with high probability. By applying the inequality recursively for $k = 0, \ldots, T$, we obtain (9). Please see [1] for the complete proof. □

Algorithm 3: ICAPM for SVM

Choose an arbitary initial point $\phi_0 \in \Re^n$, $b_0 \in \Re$;
for $k = 0, 1, 2, \ldots$ **do**
(1) Gradient discent:

$$\phi_{k+1/2} = \phi_k - \frac{1}{k+1}\phi_k\ ; \tag{11}$$

(2) Sample M training points $x_{\omega_{k,1}}, \ldots, x_{\omega_{k,M}}$ and their labels $y_{\omega_{k,1}}, \ldots, y_{\omega_{k,M}}$ from the total m points;
(3) Take the average of random projections:

$$\Delta_i = \frac{\max\{0, 1 - y_{\omega_{k,i}}(\phi'_{k+1/2}x_{\omega_{k,i}} + b_k)\}}{1 + \|x_{\omega_{k,i}}\|^2}$$

$$\phi_{k+1} = \phi_{k+1/2} + \frac{1}{M}\sum_{i=1}^{M}\Delta_i y_{\omega_{k,i}} x_{\omega_{k,i}} \tag{12}$$

$$b_{k+1} = b_k + \frac{1}{M}\sum_{i=1}^{M}\Delta_i y_{\omega_{k,i}}\ ;$$

From the preceding result, we note that the probability of satisfying the error bound is large when M_k is large. This suggests that taking averages of random projection indeed improves the stability. For a given T, we may choose M_k to be on the order of $\min\{T, m\}$ and make the probability arbitrarily close to 1. This suggests that, in order to ensure $\|x_T - x^*\|^2 \leq O(1/T)$ with probability close to 1, we need $O(T^2)$ random projections.

5. APPLICATIONS FOR ICAPM

5.1. SVM in Online Learning

Consider a classification problem in which the training points are generated online and/or the total number of training points is huge. In these cases, not all the training points are "visible" to the computer processor at once, therefore fast online algorithms are needed. Existing popular methods for large-scale SVM include the stochastic gradient descent (see [10, 11]) and the stochastic dual coordinate ascent (see [10, 11]) . Both of them apply to an unconstrained penalized optimization problem, which involves a regularization parameter λ. For real-time solution to the SVM problem, tuning λ can be expensive.

In our experiment, we consider the constrained formulation of SVM:

$$\min_{\phi,b}\ \frac{1}{2}\|\phi\|^2, \quad \text{s.t. } y_i(\phi' x_i + b) \geq 1,\ \ \forall i = 1, 2, \ldots, m. \tag{10}$$

This SVM formulation fits our general problem (1). By applying ICAMP to this SVM problem, we are able to solve this SVM problem online by sampling one or several data points (x_i, y_i) at each iteration. The implementation of ICAPM for the SVM problem (10) is given by Algorithm 3.

5.2. Network Consensus Problem

Consider a network with m agents (*nodes* in the *graph*) [12, 13] and some pair-wise links connecting the agents (*edges* in the graph).

Algorithm 4: ICAPM for Network Consensus Problem

Choose arbitary initial points $x^{(1)}, \ldots, x^{(m)} \in \Re^n$;
for $k = 0, 1, 2, \ldots$ **do**

(1) Local gradient descent for $i = 1, 2, \ldots, m$:

$$x_{k+1/2}^{(i)} = x_k^{(i)} - \frac{1}{k+1} \widetilde{\nabla} f_i(x_k^{(i)}) \; ; \tag{15}$$

(2) Sample M edges $e_i, i = 1, 2, \ldots, M$;
(3) For each sampled edge e_i that connects nodes a_i, b_i, take the average of $x^{(a_i)}$ and $x^{(b_i)}$:

$$\begin{aligned} x_{k+1}^{(a_i)} &= \frac{1}{2M}((2M-1)x_{k+1/2}^{(a_i)} + x_{k+1}^{(b_i)}) \; ; \\ x_{k+1}^{(b_i)} &= \frac{1}{2M}(x_{k+1/2}^{(a_i)} + (2M-1)x_{k+1}^{(b_i)}) \; ; \end{aligned} \tag{16}$$

We want to solve the following decentralized network optimization problem

$$\min_x \sum_{i=1}^m f_i(x), \tag{13}$$

where every agent i only knows about its local f_i. In order words, we want to obtain a *consensus* optimal solution for problem (13) by using a distributed procedure. The challenge is that we don't have a central server to enforce the consensus. Instead, we can leverage the private communication channels among agents, i.e., the links or *edges* between pairs of agents.

Motivated by the network connectivity, we rewrite (13) as:

$$\min_{x^{(1)}, x^{(2)}, \ldots, x^{(m)}} \left\{ \sum_{i=1}^m f_i(x^{(i)}) \right\}, \quad \text{s.t. } x^{(1)} = x^{(2)} = \cdots = x^{(m)}, \tag{14}$$

where $x^{(i)}$ is the local copy of the solution at the i-th agent. Given a network of nodes and edges, we obtain an optimization problem in which the objective functions are defined on the nodes while the constraints are defined on the edges. We may solve this problem by using ICPM or ICAPM. When applying ICAMP, we may interpret the stochastic gradient as agents individually making their local updates, and we may interpret the random projection as neighboring agents exchanging information. As long as the network is connected, the ICAMP is guaranteed to converge to a consensus optimal solution. The methods considered in [12, 13] also share the same spirit, while the proposed ICAPM is more general mathematically. The implementation of ICAPM for problem (13) is given by Algorithm 4.

6. EXPERIMENT RESULTS

6.1. Experiment for SVM

We generated linear separable training points and labels randomly (according to a Guassian distribution) with $x_i \in \Re^{50}$ and $m = 10^4$. Then we use Algorithm 3 to learn the optimal classifier. Here the decision variables are ϕ and b, so we quantify the optimization error by $\|x_k - x^*\|^2 := \|\phi_k - \phi^*\|^2 + |b_k - b^*|^2$.

The simulation results are given by Figure 2. We clearly observe that: ICAPM reduces the variance and keeps the iterates closer to the optimal point with high probability. Using a constant is preferable for the initial steps, while increasing M_k produces more accurate solution for large k.

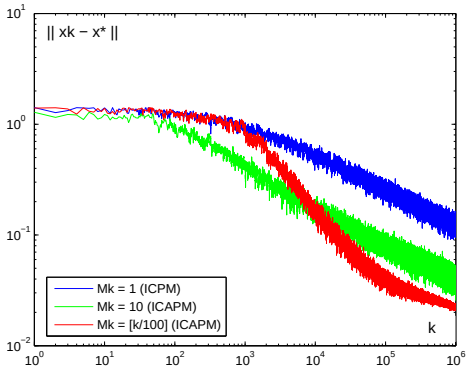

Fig. 2. ICAPM and ICPM for SVM

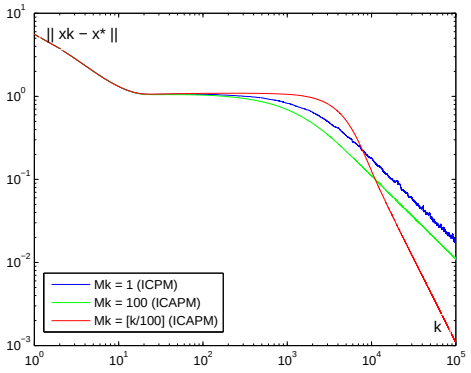

Fig. 3. ICAPM and ICPM for Network Consensus Problem

6.2. Experiment for Network Consensus Problem

Let the objective functions be $f_i(x) = \|x - p_i\|^2$, where p_i is a local knowledge only available to the i-th agent. In our experiment, we generate p_i randomly (Uniform Distribution in $(0, 1)^m$), and let $m = 10^4$, $n = 50$. The simulation results are illustrated in Figure 3, where $\sum_{i=1}^m \|x_k^{(i)} - x^*\|^2$ is the optimization error per iteration. It suggests that ICAMP is a decentralized algorithm that converges to the consensus optimal solution.

7. CONCLUSION

We proposed a stochastic optimization method for problems involving a large number of constraints, namely the ICAPM. It involves stochastic gradient descent and averaging over random projections. We prove the almost convergence of ICAPM in Theorem 1. We analyze the stability and give a high-probability error bound for ICAPM in Theorem 2 and Theorem 3, respectively. These results suggest that we can control the precision/stability of ICAPM by adjusting the number of sampling constraints per iteration.

8. REFERENCES

[1] J. Liu, Y. Gu, and M. Wang. "Averaging Randomized Projection: Online Algorithm for Convex Optimization with Many Constraints.", *Technique report*, Tsinghua University, 2014.

[2] Le Cun, Leon Bottou Yann, and L. Bottou. "Large scale online learning." *Advances in neural information processing systems* 16 (2004): 217.

[3] O. Bousquet and L. Bottou. "The tradeoffs of large scale learning." *Advances in neural information processing systems.* 2008.

[4] L. Bottou. "Large-scale machine learning with stochastic gradient descent." *Proceedings of COMPSTAT'2010.* Physica-Verlag HD, 2010. 177-186.

[5] S. Sundhar Ram, A. Nedic, and V. V. Veeravalli. "Incremental stochastic subgradient algorithms for convex optimization." *SIAM Journal on Optimization* 20.2 (2009): 691-717.

[6] A. Nedic and D. P. Bertsekas. "Incremental subgradient methods for nondifferentiable optimization[J]." *SIAM Journal on Optimization,* 2001, 12(1): 109-138.

[7] A. Nedic and D. P. Bertsekas. "Convergence rate of incremental subgradient algorithms." *Stochastic optimization: algorithms and applications.* Springer US, 2001. 223-264.

[8] M. Wang and D. P. Bertsekas. "Incremental constraint projection-proximal methods for nonsmooth convex optimization." *Technical report, MIT,* 2013.

[9] M. Wang and D. P. Bertsekas. "Incremental constraint projection methods for variational inequalities." *Mathematical Programming* (2014): 1-43.

[10] S. Shalev-Shwartz and T. Zhang. "Stochastic dual coordinate ascent methods for regularized loss." *The Journal of Machine Learning Research* 14.1 (2013): 567-599.

[11] S. Shalev-Shwartz and A. Tewari. "Stochastic methods for l 1-regularized loss minimization." *The Journal of Machine Learning Research* 12 (2011): 1865-1892.

[12] A. Nedic and A. Ozdaglar. "Distributed subgradient methods for multi-agent optimization." *Automatic Control*, IEEE Transactions on 54.1 (2009): 48-61.

[13] K. Yuan, Q. Ling, and W. Yin. "On the Convergence of Decentralized Gradient Descent." *arXiv preprint arXiv:1310.7063* (2013).

[14] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge university press, 2009.

[15] J. Duchi, E. Hazan, and Y. Singer. "Adaptive subgradient methods for online learning and stochastic optimization." *The Journal of Machine Learning Research* 12 (2011): 2121-2159.

[16] C. Hu, W. Pan, and J. T. Kwok. "Accelerated gradient methods for stochastic optimization and online learning." *Advances in Neural Information Processing Systems.* 2009.

[17] R. Jin, T. Yang, and S. Zhu. "A New Analysis of Compressive Sensing by Stochastic Proximal Gradient Descent." *arXiv preprint arXiv:1304.4680* (2013).

[18] O. Jeromin, M. S. Pattichis, and V. D. Calhoun. "Optimal compressed sensing reconstructions of fMRI using 2D deterministic and stochastic sampling geometries." *Biomedical engineering online* 11.1 (2012): 25.

[19] V. Forcen, J. Emilio, A. A. Rodriguez, and J. Garcia-Frias. "Compressive sensing detection of stochastic signals." *Information Sciences and Systems*, 2008. CISS 2008. 42nd Annual Conference on. IEEE, 2008.

[20] P. L. Combettes. "The convex feasibility problem in image recovery." *Advances in imaging and electron physics* 95.155-270 (1996): 25.

[21] P. L. Combettes and P. Bondon. "Hard-constrained inconsistent signal feasibility problems." *Signal Processing, IEEE Transactions on* 47.9 (1999): 2460-2468.

[22] E. Candes and J. Romberg. "Signal recovery from random projections." *Proc. SPIE.* Vol. 5674. 2005.