Convergence of Gradient Descent for a Class of Nonlinear Regulators

Justin Le, *Student Member, IEEE,* and Pushkin Kachroo, *Senior Member, IEEE*

***Abstract*—We derive an explicit upper bound on the error of estimates produced by gradient descent for the nonlinear regulator problem. We derive a learning rate and a modified descent step for achieving this bound. Our results show that the estimation error converges linearly to a neighborhood whose radius can be computed with only limited side information. We show the relationship between this radius and the dimension of the optimal solution.**

***Keywords*—nonlinear dynamical systems, numerical analysis**

I. Introduction

The gradient descent method, or simply *gradient descent*, has been proposed and studied for estimating solutions to two-point nonlinear boundary-value problems, namely those that arise when applying the calculus of variations to the optimal control problem [1] [2]. In this work, we study the convergence behavior of gradient descent for solving a general class of such problems. It is general in the sense that it does not presume any form for the differential equations involved, and it only requires that the cost functional be quadratic in the control function. Additionally, we show how our analyses can guide the design process in optimal control engineering.

A. Problem statement

Let $x(t), p(t) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$ at any time $t \in [t_0, t_f] \subset \mathbb{R}$ and $\mathcal{U}$ be the set of measurable functions from $[t_0, t_f]$ to $\mathbb{R}^m$. In this work, we consider the problem

$$\begin{aligned} \min_{u(t)\in\mathcal{U}} \quad & \mathcal{H} = \mathfrak{L}(t, x(t), u(t), p(t)) + \frac{1}{2}u^T(t)Ru(t) \qquad (1) \\ s.t. \quad & \dot{x}(t) = f(t, x(t), u(t), p(t)) \\ & \dot{p}(t) = g(t, x(t), u(t), p(t)) \\ & x(t_0) = x_{t_0} \\ & p(t_f) = p_{t_f}(x(t_f)). \end{aligned}$$

For brevity, we herein omit the dependent variables of each function when referring to them. The functions $\mathfrak{L}$, f, and g may be nonlinear in any argument, with the only assumption being that $\mathfrak{L}$ is differentiable w.r.t u. The matrix R is diagonal and positive definite. Although x_{t_0} is constant, p_{t_f} may be a nonlinear function of $x(t_f)$. We assume knowledge of $\mathfrak{L}$, f, g, R, x_{t_0}, and p_{t_f}.

J. Le and P. Kachroo are with the Department of Electrical and Computer Engineering, University of Nevada, Las Vegas. E-mail: justin.le@unlv.edu.
Uploaded May 1, 2017.

An example of Problem 1 arises frequently in optimal control theory, namely the two-point boundary-value problem produced by using the calculus of variations to solve the following:

$$\begin{aligned} \min_{u(t)\in\mathcal{U}} \quad & J = \int_{t_0}^{t_f} \mathfrak{L}(t, x(t)) + \frac{1}{2}u^T(t)Ru(t)dt \qquad (2) \\ s.t. \quad & \dot{x}(t) = f(t, x(t), u(t)) \\ & x(t_0) = c_1. \end{aligned}$$

The result, known as Pontryagin's minimum principle, requires the solution to Problem 2 to satisfy the equations of Problem 1 with the following conditions:

$$\begin{aligned} g &= -\frac{\partial \mathcal{H}}{\partial x} \\ p_{t_f} &= 0. \end{aligned}$$

If we further impose

$$\mathfrak{L} = \frac{1}{2}x^T Q x + p^T f,$$

then we refer to it as a regulator problem with nonlinear dynamics, or simply *nonlinear regulator (problem)*. With a slight abuse of convention, we refer to the more general Problem 1 with the same name.

More general cases of Problem 1 have been studied for applications as diverse as shape optimization for medical devices and oil reservoir management [3]. Although these applications may potentially benefit from our results, we defer the study to a future article. These applications are mentioned here to emphasize that Problem 1 is indeed more general than the nonlinear regulator of optimal control theory.

The solution to Problem 1 can only be obtained in closed form under practically restrictive circumstances. In most cases of interest, the combination of nonlinearities and split boundary conditions necessitates the use of numerical methods [1] [2]. In the interest of scaling the numerical solution to the case of large n and m, we consider here the method of gradient descent, as it achieves a dimension-free convergence rate (i.e., the bound on the error of estimates produced by gradient descent does not depend on the dimensionality m of the solution) [4] [5].

Although gradient descent has been proposed for solving Problem 1 and studied in great detail in the literature, previous works do not provide explicit formulas for describing its convergence behavior under some beneficial conditions that may be available in practice, such as side information regarding the size of the optimal solution. Another common circumstance that can be exploited is the ability to choose R with entries of arbitrarily large size, as it has been shown that the solution is only affected by the relative (and not the absolute) weighting of R w.r.t. Q [6]. More importantly, previous works provide little insight into how the convergence behavior should guide the decisions of a control system engineer. In this work, we perform such analyses, and we do so under conditions that arise naturally in practice.

B. Our contributions

For estimating the solution of Problem 1 using a modified gradient descent method, our contributions are three-fold:

- we derive an explicit upper bound on estimation error and a learning rate for achieving this bound, both of which depend mainly on the spectrum of R;
- we prove that the estimation error converges linearly to a neighborhood whose radius can be computed with only limited side information;
- we show the relationship between this radius and the dimension of the optimal solution.

The first item can guide the design of a cost function when applying our results to problems in optimal control. The second item connects our results to the known convergence rates of gradient descent and its variants, as well as to recent results in the study of linear inverse problems. The last item can guide the design of optimal controllers when the designer has influence over the number of input signals to be applied. We further discuss these benefits after presenting our main results.

The remainder of the article is organized as follows. First, we present the tools needed for our analysis, followed by the main theorem. Then, we discuss some benefits and shortcomings of our results. Finally, we discuss these benefits in the context of optimal control.

C. Notation

For variables in the problem statement and algorithm, we adopt a naming convention similar to that of the literature on optimal control theory. The Frobenius norm and ℓ_2-norm are denoted by $\|\cdot\|_F$ and $\|\cdot\|_2$, respectively. The Euclidean sphere in $\mathbb{R}^n$ is denoted by $\mathbb{S}^{n-1}$. The trace of a matrix is denoted by $\mathrm{Tr}(\cdot)$. The smallest and largest eigenvalues of a matrix M are denoted by $\lambda_{min}(M)$ and $\lambda_{max}(M)$, respectively. The value of variable u at iteration i is denoted by $u^{(i)}$. We denote the sets $\{r \in \mathbb{R}^m \mid r \geq 0\}$, $\{r \in \mathbb{R} \mid r > 0\}$, and $\{n \in \mathbb{N} \mid n > 0\}$ by the symbols $\mathbb{R}_+$, $\mathbb{R}_{++}$, and $\mathbb{N}_+$, respectively.

II. MAIN RESULTS

A. Technical overview

The proof of the main theorem shows that, with some arithmetic, the estimation error can be expressed as a sum of two terms, the *leading* and the *residual*. The leading term is bounded with the aid of Lemma II.2, and the residual is bounded with Lemma II.3. Both lemmas follow straightforwardly from Def. II.1. Meanwhile, Lemma II.1 provides an intuitive sufficiency check for Def. II.1 to be satisfied. We first present these tools before embarking on the main proof.

B. Tools

Definition II.1 ((δ, γ)-restrictedness)**.** *A matrix $M \in \mathbb{R}^{m_1 \times m_2}$ is said to be (δ, γ)-restricted over $V \subseteq \mathbb{R}^{m_2}$ if there exists $\gamma \in \mathbb{R}_+$ such that*

$$\left| \|Mv\|_2^2 - \delta \|v\|_2^2 \right| \leq \gamma,$$

for some $\delta \in \mathbb{R}_{++}$ dependent only on M and for all $v \in V$.

As δ is only dependent on M, the above definition allows the possibility of restricting M w.r.t. the same δ for different choices of V. We employ this observation in a later result.

The following lemma gives a sufficient condition for a matrix to be (δ, γ)-restricted.

Lemma II.1. *Suppose $m_1 = m_2 \equiv m$ in Def. II.1. If the eigenvalues of $M^T M$ are bounded above and below by $\delta + 1$ and $\delta - 1$, respectively, then M is (δ, γ)-restricted over $\mathbb{R}^m$.*

Proof: The result follows from the variational characterization of eigenvalues for Hermitian matrices [7]. Since $M^T M$ is Hermitian, letting $\gamma = \|v\|_2^2$ in Def. II.1 gives

$$v^T M^T M v \leq \lambda_{max}(M^T M) v^T v = (\delta + 1)\gamma,$$

for all $v \in \mathbb{R}^m$, which gives rise to one inequality in Def. II.1. The other inequality follows from a similar argument using $\lambda_{min}(M^T M)$. ■

Although Lemma II.1 does not require M to be Hermitian, we note that if $M^T M$ is diagonal, then M would indeed be Hermitian (and diagonal), a fact which will later be used.

The next lemma is a convenient restatement of Def. II.1 which will be crucial to the main proof. Along with Def. II.1, it plays a role in this article similar to the role played by Gordon's Escape Through the Mesh lemma [8] in a proof that appears in [9], from which we take inspiration for the main strategy in this article.

Lemma II.2. *Define $\mathcal{V}_- = \{v_- = v_1 - v_2 \mid v_1, v_2 \in \mathbb{S}^{m-1}\}$ and $\mathcal{V}_+ = \{v_+ = v_1 + v_2 \mid v_1, v_2 \in \mathbb{S}^{m-1}\}$. If $M \in \mathbb{R}^{m \times m}$ is (δ, γ_-)-restricted over $\mathcal{V}_-$, then for all v_-,*

$$\frac{1}{\delta} \|Mv_-\|_2^2 \leq \|v_-\|_2^2 + \frac{\gamma_-}{\delta}.$$

Furthermore, if M is (δ, γ_+)-restricted over $\mathcal{V}_+$, then for all v_+,

$$\frac{1}{\delta} \|Mv_+\|_2^2 \geq \max\left\{0,\ \|v_+\|_2^2 - \frac{\gamma_+}{\delta}\right\}.$$

Proof: The first statement follows directly from Def. II.1. The second also follows directly by noting that the left-hand side can only be positive and is hence the $\max$ between 0 and a potentially negative number. ■

It was not necessary to introduce the sets $\mathcal{V}_-$ and $\mathcal{V}_+$ in the previous lemma, but it will greatly facilitate later analysis.

The next lemma will allow us to determine the neighborhood in which the estimation error will lie after sufficiently many iterations. As mentioned, we refer to this neighborhood as the residual of the error.

Lemma II.3. *If $M \in \mathbb{R}^{m\times m}$ is diagonal, positive semidefinite, and (δ,γ)-restricted over $\mathcal{V} \subseteq \mathbb{R}^m$, then*

$$\|M^2 v\|_2 \le \sqrt{(\delta+1)\operatorname{Tr}(M^2)}\|v\|_2,$$

for all $v \in \mathcal{V}$.

Proof:

$$\begin{aligned}\|M^2 v\|_2 &\le \sqrt{\|M\|_F^2\|Mv\|_2^2}\\ &= \sqrt{\operatorname{Tr}(M^2)}\|Mv\|_2.\end{aligned}$$

Applying the definition of (δ,γ)-restrictedness to the last expression gives the result. ■

We introduce a modification to the update step of Alg. A-B which is needed in the main proof. (The algorithm is detailed in the appendix.) We motivate this decision by mentioning that the $\nabla_u \mathfrak{L}$ term in the update step may change on each iteration, as the integration steps of Alg. A-B may (and most likely will) alter the value of any variable on which it depends. As a consequence, the norm of this term may fluctuate between iterations, and we later see that the residual follows this fluctuation. To remove this iteration-dependence of the norm, we take only the direction of this term on each step:

$$\begin{aligned}u^{(i+1)} &= u^{(i)} - \mu\left(Ru^{(i)} + \xi^{(i)}\right), \qquad (3)\\ \xi^{(i)} &= \eta \frac{\nabla_u \mathfrak{L}}{\|\nabla_u \mathfrak{L}\|_2}\Big|_{u^{(i)}},\end{aligned}$$

and fix the magnitude by $\eta \in [0,\infty)$. In a sense, η serves as a parameter for tuning the "contribution" of this term to the update. By setting η to be the norm in the denominator, we recover the original update step. As we will see, η can also be used to tune the residual. A similar normalization factor to Eq. 7 has been considered in [10].

We now prove the main theorem.

Theorem II.1 (Convergence)**.** *Let $u^* \in \mathbb{R}^m$ denote the optimal solution to Problem 1 at any point in time. Let $R \in \mathbb{R}^{m\times m}$ have $\lambda_{min}(R) \ge \delta - 1$ and $\lambda_{max}(R) \le \delta + 1$ for some $\delta \in (2,\infty)$. Using $\mu = 1/\delta$ and the update step in Eq. 7, gradient descent obeys*

$$\begin{aligned}\|u^{(i)} - u^*\|_2 \le &\left(\frac{2}{\delta}\right)^i \|u^{(0)} - u^*\|_2\\ &+ \mu\left(\frac{1-\left(\frac{2}{\delta}\right)^i}{1-\frac{2}{\delta}}\right)\left(\sqrt{(\delta+1)\operatorname{Tr}(R)}\|u^*\|_2 + \eta\right), \qquad (4)\end{aligned}$$

for any number of iterations $i \in \mathbb{N}_+$ and initialized at any point $u^{(0)} \in \mathbb{R}^m$.

Proof:

$$\begin{aligned}&\|u^{(i+1)} - u^*\|_2\\ &= \|u^{(i)} - \mu(Ru^{(i)} + \xi^{(i)}) - u^*\|_2 \qquad (5)\\ &= \|(I-\mu R)(u^{(i)} - u^*) - \mu(Ru^* + \xi^{(i)})\|_2\\ &\le \|(I-\mu R)(u^{(i)} - u^*)\|_2 + \mu\left(\|Ru^*\|_2 + \|\xi^{(i)}\|_2\right)\\ &= \sup_{v_1 \in \mathbb{S}^{m-1}} v_1^T(I-\mu R)(u^{(i)} - u^*) + \mu\left(\|Ru^*\|_2 + \eta\right), \qquad (6)\end{aligned}$$

where Eq. 5 follows by definition of the update step, and Eq. 6 follows from Lemma A.1 and the definition of $\xi^{(i)}$. Again, the two terms of Eq. 6 are referred to as the leading and the residual.

We now proceed to bound the leading term. First, we observe that it can be expressed as a product involving two unit vectors by normalizing the error:

$$\begin{aligned}&v_1^T(I-\mu R)(u^{(i)} - u^*)\\ &\quad = \left(v_1^T(I-\mu R)\frac{u^{(i)} - u^*}{\|u^{(i)} - u^*\|_2}\right)\|u^{(i)} - u^*\|_2.\end{aligned}$$

To simplify the presentation, define

$$v_2 = \frac{u^{(i)} - u^*}{\|u^{(i)} - u^*\|_2}.$$

Then, note that

$$\begin{aligned}v_1^T(I-\mu R)v_2 = &\frac{1}{4}[(v_1+v_2)^T(I-\mu R)(v_1+v_2)\\ &- (v_1-v_2)^T(I-\mu R)(v_1-v_2)]\\ = &\frac{1}{4}\left[\|v_1+v_2\|_2^2 - \mu\|R^{1/2}(v_1+v_2)\|_2^2\right]\\ &- \frac{1}{4}\left[\|v_1-v_2\|_2^2 - \mu\|R^{1/2}(v_1-v_2)\|_2^2\right]. \qquad (7)\end{aligned}$$

We may now apply Lemma II.2 using $M = R^{1/2}$, $\gamma_- = \|v_-\|_2^2$, and $\gamma_+ = \|v_+\|_2^2$. These values of γ are valid as we assume the eigenvalues of R to be bounded in the sense of Lemma II.1. With the chosen value of γ_+, and with $\delta > 2$ as in the statement of the theorem, we observe that

$$\gamma_+ - \frac{\gamma_+}{\delta} > 0. \qquad (8)$$

Thus, the second bound in Lemma II.2 becomes

$$\frac{1}{\delta}\|Mv_+\|_2^2 \ge \|v_+\|_2^2 - \frac{\gamma_+}{\delta}.$$

By setting $\mu = 1/\delta$ in Eq. 7 and applying Lemma II.2,

$$\begin{aligned} v_1^T(I-\mu R)v_2 &= \frac{1}{4}\left[\gamma_+ - \frac{1}{\delta}\|R^{1/2}v_+\|_2^2\right] \\ &\quad - \frac{1}{4}\left[\gamma_- - \frac{1}{\delta}\|R^{1/2}v_-\|_2^2\right] \\ &\leq \frac{1}{4}\left[\gamma_+ - \frac{1}{\delta}\|R^{1/2}v_+\|_2^2\right] - \frac{1}{4}\left[-\frac{\gamma_-}{\delta}\right] \\ &\leq \frac{1}{4}\left(\frac{\gamma_+ + \gamma_-}{\delta}\right) \\ &\leq \frac{2}{\delta}, \end{aligned} \tag{9}$$

where the last step follows by observing that

$$\sup_{v_-\in\mathcal{V}_-}\gamma_- = \sup_{v_+\in\mathcal{V}_+}\gamma_+ = 4. \tag{10}$$

Next, by applying Ineq. 9 to Eq. 6 and applying Lemma II.3 to the residual term, we obtain a bound for the error at iteration $i+1$ given the previous iteration i:

$$\begin{aligned} \|u^{(i+1)} - u^*\|_2 &\leq \left(\frac{2}{\delta}\right)\|u^{(i)} - u^*\|_2 \\ &\quad + \mu\left(\sqrt{(\delta+1)\operatorname{Tr}(R)}\|u^*\|_2 + \eta\right) \\ &\equiv \varrho_1\|u^{(i)} - u^*\|_2 + \mu\varrho_2, \end{aligned}$$

where we have introduced the notation in the last step in order to simplify the following recursive argument.

$$\begin{aligned} \|u^{(i)} - u^*\|_2 &\leq \varrho_1\left(\varrho_1\|u^{(i-1)} - u^*\|_2 + \mu\varrho_2\right) + \mu\varrho_2 \\ &= \varrho_1^2\|u^{(i-2)} - u^*\|_2 + \mu\left(\varrho_1+1\right)\varrho_2. \end{aligned} \tag{11}$$

After applying the recursion sufficiently many times to produce $u^{(0)}$, the error term of Eq. 11 will have a coefficient of ϱ_1^i, and the term involving ϱ_2 will have the coefficient

$$\mu\left(\sum_{j=0}^{i-1}\varrho_1^j\right) = \mu\left(\frac{1-\varrho_1^i}{1-\varrho_1}\right).$$

With the appropriate substitutions for ϱ_1 and ϱ_2, this concludes the proof of Thm. II.1. ■

III. Discussion

Thm. II.1 shows that, in the limit of iterations i, the estimates converge linearly to a residual of radius $\mu\left(\frac{1}{1-\frac{2}{\delta}}\right)\left(\sqrt{(\delta+1)\operatorname{Tr}(R)}\|u^*\|_2+\eta\right)$. This radius would fluctuate with $\xi^{(i)}$ in the absence of the proposed modification in Eq. 7. The practitioner may be tempted to reduce η in this modified update step in order to reduce the residual in Thm. II.1, but we caution that sufficiently low values of η can cause gradient descent to diverge from the minimum. Intuitively, setting η to 0, for example, effectively removes the coupling between the objective and the constraint in Problem 1, and the estimate produced by gradient descent for the resulting unconstrained problem may drastically differ from the optimal solution to the constrained problem. Determining a schedule for η remains a problem for future study.

If η is chosen reasonably, the radius of the residual can be computed given side information regarding the total expended energy $\|u^*\|_2$ required by the optimal solution (i.e., the solution's energy budget). Controllers designed for resource allocation and similar applications can often afford such side information.

The following corollary provides the number of iterations required for the estimates produced by gradient descent to converge to the residual.

Corollary III.1. *Given an error bound $\epsilon > 0$, gradient descent yields an estimate $u^{(i)}$ such that, after a number of iterations $i > \ln(\epsilon)/\ln(2/\delta)$,*

$$\begin{aligned} \|u^{(i)} - u^*\|_2 - \mu\left(\frac{1-\left(\frac{2}{\delta}\right)^i}{1-\frac{2}{\delta}}\right)\left(\sqrt{(\delta+1)\operatorname{Tr}(R)}\|u^*\|_2+\eta\right) \\ \leq \epsilon. \end{aligned} \tag{12}$$

Proof: First observe that

$$x^{1/\ln(x)} = e. \tag{13}$$

To see this, let the right-hand side be any real number and take $\ln$ of both sides. From this observation and Thm. II.1, letting $x = 2/\delta$ and raising both sides of Eq. 13 to the power of $\ln(\epsilon)$, we have the result. ■

A. Application: optimal control

We conclude this section by drawing attention to some aspects of Thm. II.1 which may benefit the design of an optimal control system. The nonlinear regulator problem produced by combining Eqs. 1 and 2 has a solution u^* with dimensionality m (at each point in time). In some situations, the practitioner has influence over the number of input signals to be used at each time instant for controlling a process described by $\dot{x}$, and hence the practitioner may select m to tradeoff between various design goals. Thm. II.1 may provide some insight into this decision.

Consider that, for sufficiently large δ and after sufficiently many iterations i, the bound in Thm. II.1 is dominated by $\sqrt{m}$. To see this, note that $\mu = 1/\delta$ and that R is a diagonal (δ,γ)-restricted matrix of size m, which leads to $\operatorname{Tr}(R) \approx \delta m$. The practitioner may choose m in accordance with the desired bound on estimation error. Additionally, a sufficiently large δ may be enforced in order to exploit this observation, since the size of entries in R can be increased arbitrarily in any problem as long as the relative weight between R and Q is preserved [6].

The practitioner may similarly have access to side information regarding $\|u^*\|_2$, as in systems whose controllers have only a fixed "budget" of input energy. If the budget is known during the design process, it can be used to compute the radius of the residual in Thm. II.1 as mentioned.

IV. Conclusion

We studied the convergence behavior of gradient descent for a general class of nonlinear programs with differential equation constraints. We showed that the estimates produced by gradient descent converge linearly to a residual whose radius can be explicitly computed given side information regarding the spectrum of the weighting matrix R in the objective and the norm of the optimal solution at all points in time. The derived bounds gave insight into the choice of learning rate, as well as the choice of the input dimensionality in the design of optimal control systems.

Appendix A
Supplement

A. The supremum definition of ℓ_2-norm

Lemma A.1. *Let $w_1 \in \mathbb{R}^n$. Then*

$$\|w_1\|_2 = \sup_{w_2 \in \mathbb{S}^{n-1}} w_2^T w_1.$$

Proof: Let θ be the smallest angle between w_1 and w_2. If $\|w_2\|_2 = 1$, then

$$w_2^T w_1 = (\|w_1\|_2 \|w_2\|_2)\, cos(\theta) \leq \|w_1\|_2,$$

for all $w_1 \in \mathbb{R}^n$. ■

B. Gradient descent for Problem 1

Initialize $u^{(0)}$
i = 1
while $\|\partial\mathcal{H}/\partial u\|_2 > \epsilon$ **do**
  Integrate $\dot{x}^{(i)}$ from t_0 to t_f
  Compute $p_{t_f}^{(i)}$ using $x_{t_f}^{(i)}$
  Integrate $\dot{p}^{(i)}$ from t_f to t_0
  $u^{(i)} = u^{(i)} - \mu\nabla_u \left(\mathfrak{L} + \frac{1}{2}u^T R u\right)\Big|_{u^{(i)}}$,
  $i = i + 1$
end while

References

[1] A. E. Bryson and Y. C. Ho, *Applied Optimal Control*. New York: Blaisdell, 1969.

[2] D. E. Kirk, *Optimal control theory; an introduction*. Prentice-Hall, Apr. 1970.

[3] L. T. Biegler, O. Ghattas, M. Heinkenschloss, and B. van Bloemen Waanders, *Large-Scale PDE-Constrained Optimization: An Introduction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 3–13. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-55508-4_1

[4] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed. Springer Publishing Company, Incorporated, 2014.

[5] S. Bubeck, "Convex optimization: Algorithms and complexity," *Found. Trends Mach. Learn.*, vol. 8, no. 3-4, pp. 231–357, Nov. 2015. [Online]. Available: http://dx.doi.org/10.1561/2200000050

[6] R. F. Stengel, *Stochastic Optimal Control: Theory and Application*. New York, NY, USA: John Wiley & Sons, Inc., 1986.

[7] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York, NY, USA: Cambridge University Press, 1986.

[8] Y. Gordon, *On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1988, pp. 84–106. [Online]. Available: http://dx.doi.org/10.1007/BFb0081737

[9] S. Oymak, B. Recht, and M. Soltanolkotabi, "Sharp time-data tradeoffs for linear inverse problems," *CoRR*, vol. abs/1507.04793, 2015. [Online]. Available: http://arxiv.org/abs/1507.04793

[10] A. N. Iusem, "On the convergence properties of the projected gradient method for convex optimization," *Computational & Applied Mathematics*, vol. 22, pp. 37 – 52, 00 2003.