Nonconvex optimization

R-local Minimizers, Global Optimality, and Run-and-Inspect Methods

Yifan Chen Yuejiao Sun Wotao Yin

Overview

Introduction

1D risk function

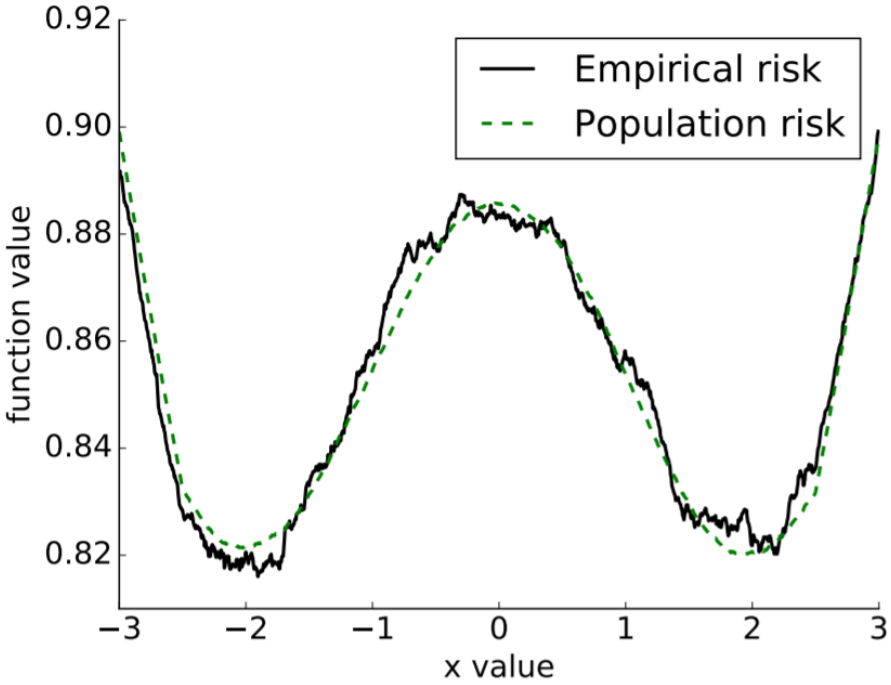


Figure 1: Risk functions based on 5000 samples (empirical) vs population (sample $\to \infty$).[Zhang et al., 2017]

A motivating example

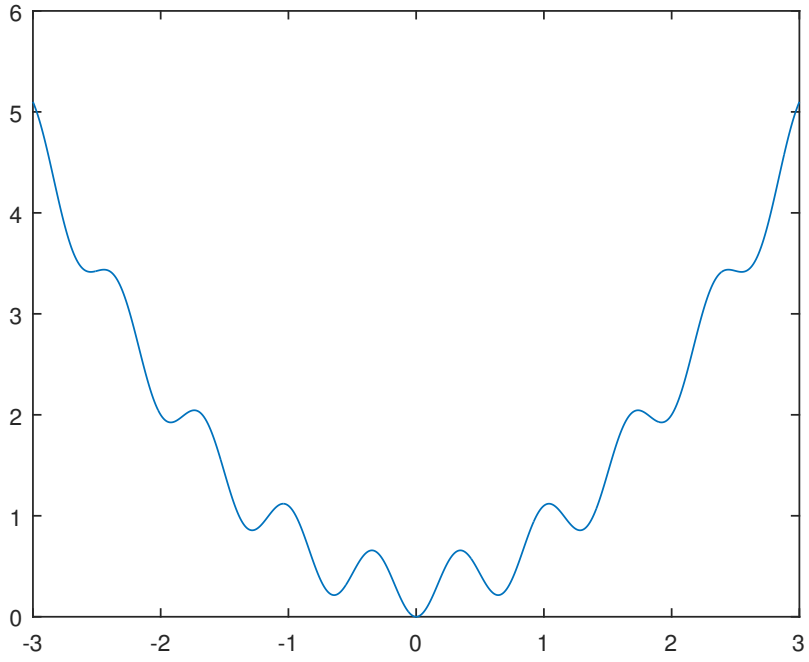


Figure 2: The graph of $F(x) = \frac{x^2}{2} + 0.3\sin\left(3\pi(x - \frac{1}{6})\right) + 0.3$

- convex + perturbation $\Rightarrow$ nonconvex + local min
- Gradient Descent may get stuck in a local minimizer.
- how to escape from local minimizers?

A motivating example

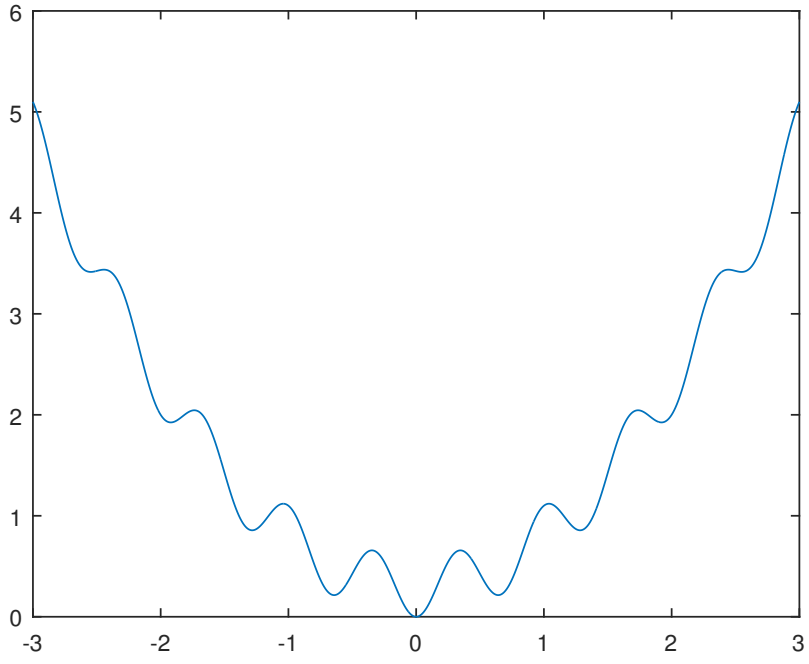


Figure 3: The graph of $F(x) = \frac{x^2}{2} + 0.3\sin\left(3\pi(x - \frac{1}{6})\right) + 0.3$

- "Run": Gradient Descent converges to a stationary point $\bar{x}$;
- "Inspect": look for $y \in [\bar{x} - R, \bar{x} + R]$ such that $F(y) < F(\bar{x})$;

If $R > \min\{2\sqrt{a}, \frac{2}{b}\}$, it will converge to the unique global minimizer.

Overview

- problem: minimize $F(\mathbf{x}) =$ convex/smooth $+$ perturbation.
- approach: find an R-local minimizer.
- guarantee: approximate global minimizers, dimension-polynomial complexity.

Theoretical Analysis

Model assumption

Implicit "convex/smooth + perturbation" decomposition

$$F(\mathbf{x}) = f(\mathbf{x}) + r(\mathbf{x}).$$

- $f(\mathbf{x})$ is differentiable, and $\nabla f(\mathbf{x})$ is L-Lipschitz continuous.
- $|r(\mathbf{x}) - r(\mathbf{y})| \leq \alpha\|\mathbf{x} - \mathbf{y}\| + 2\beta$, e.g. $\ell_{1/2}$, MCP,...

(blockwise) R-local minimizer

Definition

$\bar{\mathbf{x}}$ is called a standard R-local minimizer of F if it satisfies

$$F(\bar{\mathbf{x}}) = \min_{\mathbf{x} \in B(\bar{\mathbf{x}}, R)} F(\mathbf{x}).$$

Definition

$\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_s)$ is called a blockwise **R**-local minimizer of F if it satisfies

$$F(\bar{x}_i, \bar{\mathbf{x}}_{-i}) = \min_{x_i \in B(\bar{x}_i, R_i)} F(x_i, \bar{\mathbf{x}}_{-i}), \quad 1 \leq i \leq s,$$

where $F(\bar{\mathbf{x}}) = F(\bar{x}_i, \bar{\mathbf{x}}_{-i})$.

Blockwise version is proposed for high dimensional problems.

Analysis results

Recall : $F(\mathbf{x}) = f(\mathbf{x}) + r(\mathbf{x})$

For F = smooth + perturbation,

$$\{ \text{ (blockwise) R-local min } \} \subseteq \{ \mathbf{x} : \|\nabla f(\mathbf{x})\| \leq \delta \}$$

For F = smooth, strongly convex + perturbation,

- $\{ \mathbf{x} : \|\nabla f(\mathbf{x})\| \leq \delta \} \subset \{$ approx global min $\}$
- $\{$ (blockwise) R-local min $\} \subseteq \{$ approx global min $\}$

$\{$ (Blockswise) R-local min $\} \subseteq \{ \mathbf{x} : \|\nabla f(\mathbf{x})\| \leq \delta \}$

Theorem

If $\bar{\mathbf{x}}$ is an R-local minimizer with $R \geq 2\sqrt{\frac{\beta}{L}}$, then

$$\|\nabla f(\bar{\mathbf{x}})\| \leq \delta = \alpha + 2\sqrt{\beta L}.$$

Proof: Minimize RHS for $\mathbf{x} \in B(\bar{\mathbf{x}}, R)$,

$$0 \leq F(\mathbf{x}) - F(\bar{\mathbf{x}}) \leq 2\beta + \alpha\|\mathbf{x} - \bar{\mathbf{x}}\| + \langle \nabla f(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}} \rangle + \frac{L}{2}\|\mathbf{x} - \bar{\mathbf{x}}\|^2$$

Theorem

If $\bar{\mathbf{x}}$ is a blockwise $\mathbf{R}$-local minimizer of F with $R_i \geq 2\sqrt{\frac{\beta}{L_i}}$, then $\|\nabla f(\bar{\mathbf{x}})\| \leq \delta = \|\mathbf{v}\| := (\sum |v_i^2|)^{\frac{1}{2}}$ where $v_i := \alpha + 2\sqrt{\beta L_i}$, $1 \leq i \leq s$. Specially, we have

$$\|\nabla f(\bar{\mathbf{x}})\| \leq \delta = \sqrt{s}\left(\alpha + 2\sqrt{\beta L}\right).$$

Theorem

Suppose f is μ-strongly convex. If $\|\nabla f(\bar{\mathbf{x}})\| < \delta$, then

$$F(\bar{\mathbf{x}}) - F^* \le \frac{\delta^2 + 2\alpha\delta}{\mu} + 2\beta, \qquad \|\bar{\mathbf{x}} - \mathbf{x}^*\| \le \frac{2\delta}{\mu}.$$

- (blockwise) R-local min $\Rightarrow$ approx global min.
- R-local min with $R \ge \frac{2\delta}{\mu} = \frac{2\left(\alpha + 2\sqrt{\beta L}\right)}{\mu} \Rightarrow$ global min.
- blockwise R-local min with $R_i \ge \frac{2\left(\alpha + 2\sqrt{\beta L}\right)}{\mu} \Rightarrow$ Nash point.

Run-and-Inspect Method

Approximated R-local minimizers

A point $\bar{\mathbf{x}}$ is called a blockwise **R**-local minimizer of F *up to* $\eta = [\eta_1 \cdots \eta_s]^T \geq 0$ if it satisfies

$$F(\bar{x}_i, \bar{\mathbf{x}}_{-i}) \leq \min_{x_i \in B(\bar{x}_i, R_i)} F(x_i, \bar{\mathbf{x}}_{-i}) + \eta_i, \quad 1 \leq i \leq s;$$

Theorem

If $\bar{\mathbf{x}}$ *is a blockwise* **R***-local minimizer of* F *up to* η *for* $R_i \geq \sqrt{\frac{4\beta+2\eta_i}{L_i}}$, *then* $\|\nabla f(\bar{\mathbf{x}})\| \leq \delta \geq \|\mathbf{v}\| := (\sum |v_i^2|)^{\frac{1}{2}}$ *for* $v_i = \alpha + \sqrt{(4\beta + 2\eta_i)L_i}$, $1 \leq i \leq s$. *Specially,*

$$\nabla f(\bar{\mathbf{x}}) \leq \delta = \sqrt{s}\left(\alpha + \sqrt{(4\beta + 2\|\eta\|_\infty)L}\right).$$

Run-and-Inspect method

Alternate steps 1 and 2 below:

1. **Run** a descent algorithm to a stationary point;
2. **Inspect** the R-radius of $\bar{\mathbf{x}}$ by sampling (blockwisely), looking for a point with sufficient descent ν¿0:
   - if found, resume step 1 from that point;
   - otherwise, return $\bar{\mathbf{x}}$.

It guarantees to converge to an approx (blockwise) R-local min.

Inspection by deterministic sampling

Theorem

Assume that $f(x_i, \bar{\mathbf{x}}_{-i})$ is $\bar{L}_i$-Lipschitz continuous in the ball $B(\bar{x}_i, R_i)$. If we sample points with density $\bar{r}$ blockwisely, then $\bar{\mathbf{x}}$ is a blockwise $\mathbf{R}$-local minimizer of F up to $\eta_i = \nu + (\bar{L}_i + \alpha)\bar{r} + 2\beta$.

Theorem

When $\|\nabla_i f(\bar{x}_i, \bar{\mathbf{x}}_{-i})\| \geq \frac{\frac{L_i}{2} R_i^2 + \alpha R_i + 2\beta + \nu}{R_i - \bar{r}}$, the algorithm can escape to a better point certainly.

Complexity of Blockwise Run-and-Inspect

Parameters.

- $s = \frac{d}{d'}$, block size d', $\|x^0 - x^*\| = O(d^u)$;
- $\alpha = o(1), \beta = o(1), L_i = \Theta(1), \mu = \Theta(1)$;
- $\bar{r} = t\beta$, $R_i = R = \Theta(\sqrt{(1+t)\beta})$, $\bar{r} \leq \frac{R}{4}$. $\eta = \Theta((1+t)\beta)$.

The global optimality bound is

$$\frac{F(\bar{\mathbf{x}}) - F(\mathbf{x}^*)}{F(\mathbf{x}^0) - F(\mathbf{x}^*)} = O\left(\frac{d^{1-2u}}{d'}\left(\alpha + \sqrt{(1+t)\beta}\right)^2\right).$$

The total complexity is

$$O\left(\frac{d^{2u}}{t\beta}\frac{d}{d'}e^{O\left(d'\log\left(\sqrt{\frac{1+t}{t^2\beta}}\right)+d'\right)}\right).$$

Numerical Results

Numerical experiments

Nonconvex robust linear regression:

- Linear model: $y = \langle \beta, \mathbf{x} \rangle + \varepsilon$.
- Data points: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ with outliers.
- Tukey's bisquare loss:

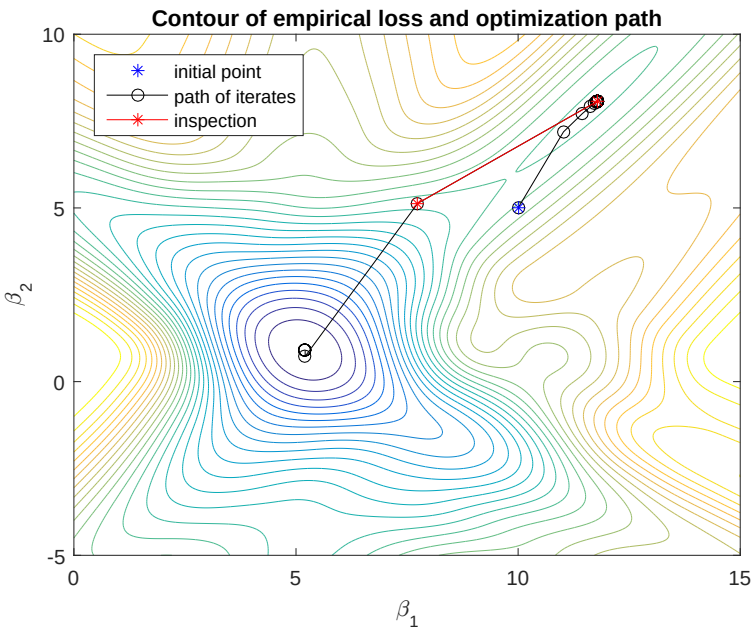
$$\rho(r) = \begin{cases} \frac{r_0^2}{6}\{1 - (1 - (r/r_0)^2)^3\}, & \text{if } |r| < r_0, \\ \frac{r_0^2}{6}, & \text{otherwise.} \end{cases}$$

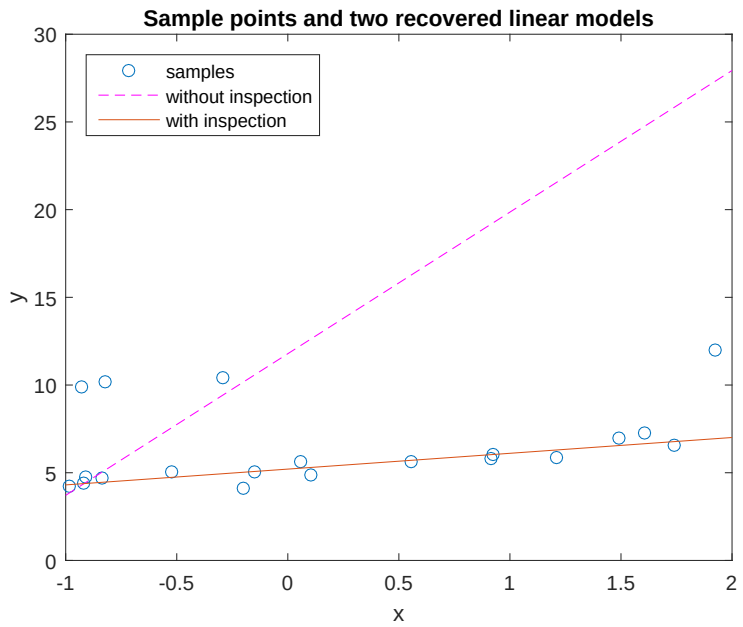
The empirical loss function based on ρ is

$$l(\beta) = \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - \langle \beta, x_i \rangle).$$

Nonconvex robust linear regression

- The test uses the model $y = 5 + x + \varepsilon$.
- Generate $x_i \sim \mathcal{N}(0, 1)$, $\varepsilon_i \sim \mathcal{N}(0, 0.5)$, $i = 1, 2, \ldots, 20$ Create 20% outliers by adding extra noise generated from $\mathcal{N}(0, 5)$.
- Use IRLS and Run-and-Inspect Method in 2D withwith $R = 5, dR = 0.5, \nu = 10^{-3}$.





Numerical experiments

Nonconvex compressed sensing:

- Given $A \in \mathbb{R}^{m\times n}$ $(m < n)$ and sparse $\mathbf{x} \in \mathbb{R}^n$, observe

$$\mathbf{b} = A\mathbf{x}.$$

- Problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; Q(\mathbf{x}) := \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2 + \lambda\|\mathbf{x}\|_{\frac{1}{2}}^{\frac{1}{2}},$$

- Coordinate descent

$$\begin{aligned} : x_j^{k+1} &= \operatorname*{argmin}_{x_j} Q(x_j, \mathbf{x}_{-j}^k) \\ &= \operatorname*{argmin}_{x_j} \; \frac{1}{2}A_j^T A_j x_j^2 + A_j^T(A\mathbf{x}^k - \mathbf{b})x_j + \lambda\sqrt{|x_j|}. \end{aligned}$$

  It has a closed-form solution.

Nonconvex compressed sensing

- Set $m = 25, 50, 100$ and $n = 2m$. GenerateA_{ij} from $\mathcal{U}(0, \frac{1}{\sqrt{m}})$ i.i.d. $\mathbf{x}$ has 10% nonzeros generated from $\mathcal{U}(0.2, 0.8)$ i.i.d. Set $b = A\mathbf{x}$.
- Apply coordinate descent and 2D inspection (CDI) with $R = 0.5$, $\Delta R = 0.05$. And compared it with standard coordinate descent (CD) and half thresholding algorithm (*half*) [Xu et al., 2012].

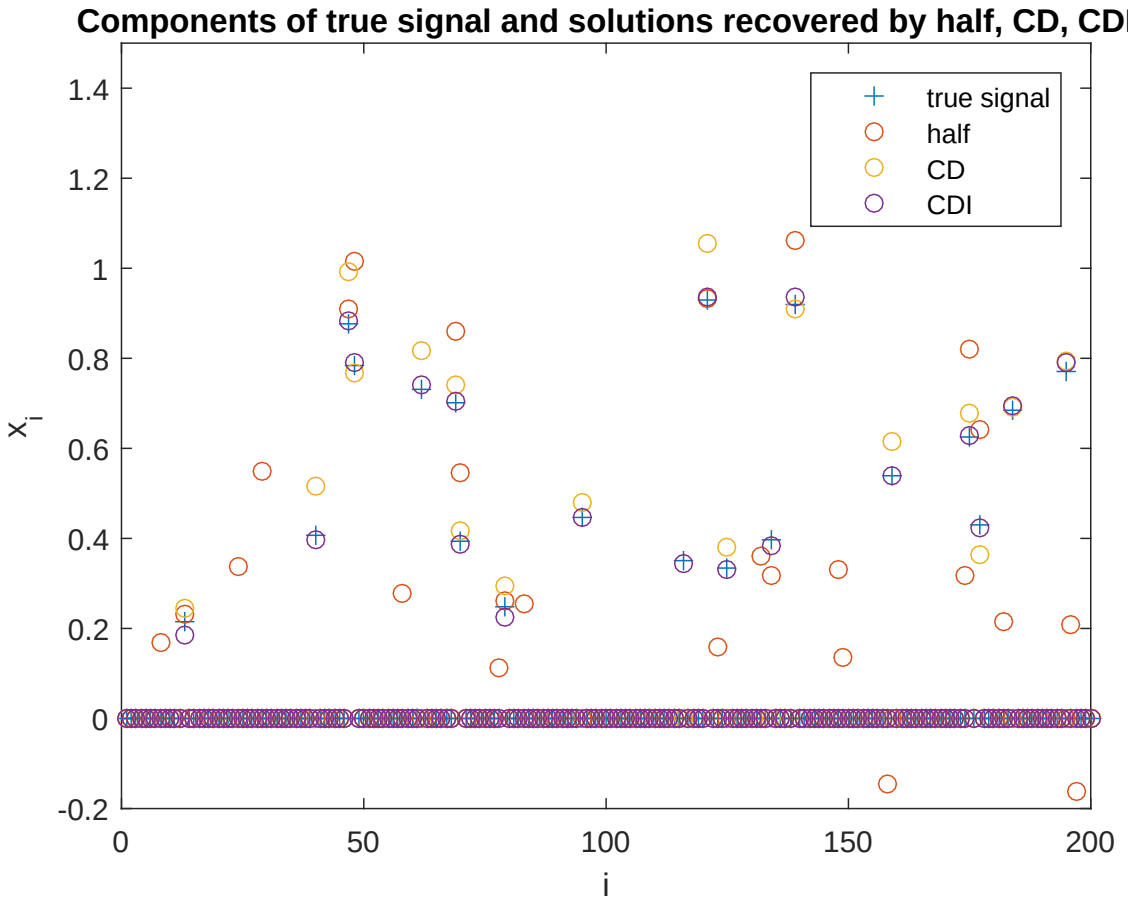
Compressed sensing

n, p	algorithm	a	b	c	ave obj
$n = 25$	*half*	47.73%	2	2	0.0365
$p = 50$	CD	62.40%	25	27	0.0272
	CDI	83.95%	65	69	0.0208
$n = 50$	*half*	46.43%	0	0	0.0736
$p = 100$	CD	76.39%	24	32	0.0443
	CDI	92.34%	57	68	0.0369
$n = 100$	*half*	44.31%	0	0	0.1622
$p = 200$	CD	85.97%	10	18	0.0795
	CDI	94.31%	54	76	0.0756

Table 1: Statistics of 100 compressed sensing problems.

1. a is the average ratio of correctly identified nonzeros to true nonzeros
2. b is the number of tests with all true nonzeros identified;
3. c is the number of tests in which the returned points yield lower objective values than that of the true signal (only model error, no algorithm error).

Nonconvex compressed sensing



In one experiment, CDI recovered all positions of nonzeros of x, while CD failed to recover x_{116}, x_{134}. The *half* algorithm just got stuck at a local minizer far from x.

Future plan

Future plan

- improve the efficiency of inspection
- applications
- how to estimate or choose parameters

References

Xu, Z., Chang, X., Xu, F., and Zhang, H. (2012). $I\_\{1/2\}$ regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on neural networks and learning systems*, 23(7):1013–1027.

Zhang, Y., Liang, P., and Charikar, M. (2017). A hitting time analysis of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1702.05575*.

Thank you!