OPT2024: 16th Annual Workshop on Optimization for Machine Learning

Nonlinear tomographic reconstruction via nonsmooth optimization

Vasileios Charisopoulos VCHARISO@UCHICAGO.EDU
Data Science Institute
University of Chicago

Rebecca Willett WILLETT@UCHICAGO.EDU
Computer Science, Statistics, Computational and Applied Mathematics, and Data Science Institute
University of Chicago

Abstract

We study iterative signal reconstruction in computed tomography (CT), wherein measurements are produced by a linear transformation of the unknown signal followed by an exponential nonlinear map. Approaches based on pre-processing the data with a log transform and then solving the resulting linear inverse problem are amenable to convex optimization methods but perform poorly for signals with high dynamic range, as in X-ray imaging of tissue with embedded metal. We show that a suitably initialized subgradient method applied to a natural nonsmooth, nonconvex loss function produces iterates that converge to the unknown signal of interest at a geometric rate under a recently proposed statistical model. Our recovery program enables faster iterative reconstruction from substantially fewer samples.

1. Introduction

Computed tomography (CT) scans are X-ray imaging procedures widely used in medicine [3], security screenings [28], non-destructive material evaluation [4], archaeology [6], and others. Several commercial CT scanners rely on the following model [20]: given a finite set of illumination angles, $\{\theta_i\}_{i=1}^m$, we collect measurements

$$y_i = I_0 \exp\left(-\langle a_i, x_\star \rangle\right), \quad \text{for } i = 1, \dots, m, \tag{1}$$

where $\{a_i\}_{i=1}^m \subset \mathbb{R}^d$ are known coefficients derived from the Radon transform [23] at angles $\{\theta_i\}_{i=1}^m$ and $x_\star \in \mathbb{R}^d$ is a vectorized representation of the unknown image. To recover $x_\star$, CT software applies a logarithmic preprocessing step to the y_i [14], resulting in the linear system

$$\langle a_i, x_\star \rangle = \hat{y}_i, \quad i = 1, \dots, m, \quad \text{where} \quad \hat{y}_i = \log\left(I_0/y_i\right). \tag{2}$$

The linear inverse problem in (2) is then solved using traditional methods, most commonly the so-called *filtered back-projection* (FBP) method. However, the logarithmic transform is numerically unstable as $y_i \to 0$, which is often the case for X-rays passing through high-density materials, and is known to lead to reconstruction artifacts [16]. This motivates researchers and practitioners to consider iterative reconstruction methods that operate directly on (1).

Recently, Fridovich-Keil et al. [13] studied iterative reconstruction from CT measurements under a Gaussian measurement model, intended to capture randomness in ray directions. They postu-

late that when the design vectors a_i in (1) are i.i.d. Gaussian, $x_\star$ can be recovered by solving

$$\widehat{x} = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2m} \sum_{i=1}^{m} \left(y_i - e^{-\langle a_i, x \rangle_+} \right)^2 . \tag{3}$$

While working with the measurements directly avoids numerical instability, the optimization problem in (3) is nonconvex; consequently, iterative methods like gradient descent are not guaranteed to converge to the solution $x_\star$. Nevertheless, Fridovich-Keil et al. [13] prove the following:

Theorem 1.1 (Informal; adapted from [13, Theorem 1]) *Suppose that $a_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ and the number of measurements m satisfies $m \gtrsim \frac{e^{c\|x_\star\|}}{\|x_\star\|^2} \cdot d$. Then solving (3) using gradient descent with stepsize η and careful initialization produces a sequence of iterates $\{x_k\}_{k \geq 1}$ satisfying*

$$\|x_k - x_\star\|^2 \leq \left(1 - \eta e^{-5\|x_\star\|}\right)^k \|x_\star\|^2 , \tag{4}$$

as long as $\eta \lesssim e^{-5\|x_\star\|}$, with high probability over the choice of design vectors.

Unfortunately, both sample and iteration complexity in Theorem 1.1 depend on $\|x_\star\|$ **exponentially**.

Our contribution. A natural question is whether the aforementioned limitations are essential or can be sidestepped by choosing a different objective function or reconstruction method. We show that using a nonsmooth loss function — namely, the least absolute deviation (ℓ_1) penalty — and optimizing it with a suitable modification of the subgradient method with Polyak step [22] yields *exponential improvements to both sample and iteration complexity*; see Table 1.

Method	Iteration complexity	Sample complexity	Reference
(**PolyakSGM**)	$O\left(\|x_\star\|^6 \log\left(\frac{\|x_\star\|}{\varepsilon}\right)\right)$	$O(\|x_\star\|^4 \cdot d)$	Theorem 2.1
Gradient descent	$O\left(e^{c_1\|x_\star\|} \log\left(\frac{\|x_\star\|}{\varepsilon}\right)\right)$	$O\left(\frac{e^{c_2\|x_\star\|}}{\|x_\star\|^2} \cdot d\right)$	Fridovich-Keil et al. [13]

Table 1: Iteration and sample complexity of iterative reconstruction methods.

1.1. Method overview

We propose optimizing the following nonsmooth (ℓ_1) penalty:

$$f(x) := \frac{1}{m} \sum_{i=1}^{m} \left| y_i - \exp(-\langle a_i, x \rangle_+) \right| . \tag{5}$$

To optimize $f(x)$ in (5), we use the subgradient method with Polyak step-size initialized at $x_0 = \mathbf{0}$, labeled as (**PolyakSGM**) below. Note that $f_\star = f(x_\star) = 0$ in the absence of noise.

(**PolyakSGM**) $$x_{k+1} = x_k - \eta \cdot \frac{f(x_k) - f_\star}{\|v_k\|^2} v_k, \quad v_k \in \partial f(x_k), \quad \text{for } k = 0, 1, \ldots$$

Here, $\partial f(x)$ denotes the so-called *Clarke* subdifferential [9] (see Appendix C). Our analysis will focus on showing that the loss f in (5) is "well-conditioned" in a neighborhood of the solution $x_\star$, a consequence of the following two regularity properties:

$$\textbf{(Lipschitz continuity):} \qquad |f(x) - f(\bar{x})| \le \mathsf{L} \|x - \bar{x}\|, \quad \text{for all } x, \bar{x} \in \mathbb{R}^d; \tag{6a}$$

$$\textbf{(Sharp growth):} \qquad f(x) - f(x_\star) \ge \mu \|x - x_\star\|, \quad \text{for all } x \text{ near } x_\star. \tag{6b}$$

It is well known (see, e.g., Goffin [17], Polyak [22]) that classical subgradient methods converge linearly for *convex* functions satisfying (6a)–(6b), with rate governed by the nonsmooth *condition number* $\kappa := \mathsf{L}/\mu$. While the loss in (5) is nonconvex, we show that it satisfies a key "aiming" condition, postulating that subgradients point to the direction of the solution $x_\star$.

$$\textbf{(Aiming)} \qquad \min_{v \in \partial f(x)} \langle v, x - x_\star \rangle \ge \mu \|x - x_\star\|, \quad \text{for all } x \in \mathcal{B}(\mathbf{0}; 3\|x_\star\|) \setminus \{0\}.$$

The (**Aiming**) inequality serves two purposes: first, it implies the sharp growth property (6b) in a neighborhood of $x_\star$ (Lemma 2.1). Moreover, it ensures iterates of (**PolyakSGM**) approach $x_\star$:

$$\|x_{k+1} - x_\star\|^2 \le \|x_k - x_\star\|^2 - \frac{\eta\mu f(x_k)}{\|v_k\|^2} \|x_k - x_\star\| < \|x_k - x_\star\|^2, \quad \text{for } \eta \text{ small enough.}$$

The focal point of our analysis is establishing (**Aiming**) and quantifying the moduli of sharp growth and Lipschitz continuity. We show that, with high probability, Gaussian designs satisfy

$$\mathsf{L} = O(1), \quad \text{and} \quad \mu = \Omega\left(\|x_\star\|^{-2}\right), \quad \text{as long as } m \gtrsim d\|x_\star\|^4. \tag{7}$$

Consequently, we establish that the conditioning of (5) is **polynomial** in $\|x_\star\|$, improving exponentially upon the results of [13], and prove **linear convergence** of (**PolyakSGM**) to $x_\star$.

1.2. Related work

Several algorithms for CT image reconstruction, including the standard FBP method, can be viewed as discrete approximations of analytical inversion formulas and are relatively inexpensive to implement, in contrast to iterative reconstruction (IR) methods; see [2, 15, 21]. While a comprehensive overview of IR methods is beyond the scope of this article, and can be found in surveys such as [2, 30], several of these methods are motivated by advances in numerical optimization: examples include the classical algebraic reconstruction technique (ART) [18], the SAGE method of Fessler and Hero [11], the ASD-POCS method of Sidky and Pan [24], and the nonconvex ADMM approach of Barber and Sidky [1]. Few of these works provide estimates on the sample and computational efficiency of the proposed methods and, when they do, the estimates are typically not adapted to CT problems. Beyond computed tomography, several works design and analyze first-order methods for signal recovery in other settings; this includes works that study the sample and computational complexity of recovering a signal from magnitude-only or quadratic measurements [26] (also known as *phase retrieval*) and measurements produced by piecewise nonlinearities such as ReLUs [12, 25], as well as recovering low-rank matrices using the Burer-Monteiro factorization [5, 19].

Notation and basic constructions. We write $\langle X, Y \rangle := \mathsf{Tr}(X^\mathsf{T} Y)$ for the Euclidean inner product and $\|X\| = \sqrt{\langle X, X \rangle}$ for the induced norm. We denote the unit sphere in d dimensions by $\mathbb{S}^{d-1}$ and the Euclidean ball centered at $\bar{x}$ and radius r by $\mathcal{B}(\bar{x}; r)$. We write $\|A\|_{\text{op}} := \sup_{x \in \mathbb{S}^{d-1}} \|Ax\|$ for the spectral norm of A. Finally, we write $A \lesssim B$ to indicate that $A \le c \cdot B$ for a constant $c > 0$.

2. Linear convergence of (PolyakSGM)

We now establish the linear convergence of (**PolyakSGM**). In our analysis, we assume that f is L-Lipschitz (6a) and satisfies (**Aiming**); quantitative estimates and formal proofs for both properties can be found in Appendix A. We first show that (**Aiming**) nearly implies local sharp growth:

Lemma 2.1 (Solvability lemma) *Suppose f is Lipschitz and (**Aiming**) holds. Then we have:*

$$f(x) - f_\star \geq \mu \min\{\|x - x_\star\|, \|x\|\}, \quad \text{for all } x \in \mathcal{B}(x_\star; \|x_\star\|). \tag{8}$$

While (8) is weaker than (6b), we show that the norm of the iterates produced by (**PolyakSGM**) stays bounded away from $\mathbf{0}$ and eventually surpasses the distance to $x_\star$, whereupon (8) reduces to sharp growth. Indeed, we have the following implication (see Lemma E.5 for a proof):

$$\|x - x_\star\|^2 \leq (1 - \gamma_0)\|x_\star\|^2 \implies \min\{\|x - x_\star\|, \|x\|\} \geq \min\left\{1, \frac{\gamma_0}{2\sqrt{1-\gamma_0}}\right\} \cdot \|x - x_\star\|. \tag{9}$$

In particular, when $\eta \leq \frac{1}{\kappa}$ the first iterate of (**PolyakSGM**) satisfies the antecedent condition in (9) for some γ_0 that depends on the design matrix. The convergence analysis proceeds as follows:

$$\begin{aligned}
\|x_{k+1} - x_\star\|^2 &= \|x_k - x_\star\|^2 - \frac{\eta f(x_k)}{\|v_k\|^2}\left(2\langle v_k, x_k - x_\star\rangle - \eta f(x_k)\right) \\
&\leq \|x_k - x_\star\|^2 - \frac{\eta f(x_k)}{\|v_k\|^2}\left(2\mu\|x_k - x_\star\| - \eta\mathsf{L}\|x_k - x_\star\|\right) \quad (\text{(Aiming)} + (6a)) \\
&\leq \|x_k - x_\star\|^2 - \frac{\eta\mu f(x_k)\|x_k - x_\star\|}{\mathsf{L}^2}, \quad (\eta \leq \mu/\mathsf{L})
\end{aligned} \tag{10}$$

showing that $\|x_{k+1} - x_\star\|^2 < \|x_k - x_\star\|^2$ unless $f(x_k) = f_\star = 0$. To deduce linear convergence, it remains to argue that $f(x_k)$ is lower bounded by $c \cdot \|x_k - x_\star\|$. Indeed, we consider:

- $\|x_k - x_\star\| > \frac{1}{2}\|x_\star\|$; while this is true, we use Lemma 2.1 and (9) to deduce
$$f(x_k) \geq (\mu\gamma_0/4)\|x_k - x_\star\|.$$
- $\|x_k - x_\star\| \leq \frac{1}{2}\|x_\star\|$; by Claim 6, $\|x_k\| \geq \frac{1}{2}\|x_\star\|$. Lemma 2.1 then implies
$$f(x_k) \geq \mu\min\{\|x_k - x_\star\|, \|x_k\|\} \geq \mu\min\left\{\|x_k - x_\star\|, \frac{1}{2}\|x_\star\|\right\} = \mu\|x_k - x_\star\|.$$

We can also guarantee that **(i)** iterates enter the ball $\mathcal{B}(x_\star; \frac{1}{2}\|x_\star\|)$ within $O(\kappa^3)$ iterations and **(ii)** never escape that ball. Our main result, whose full proof can be found in Appendix B, is as follows:

Theorem 2.1 *Let $a_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ and $m \gtrsim d \cdot \|x_\star\|^4$. Then with probability at least $1 - 4e^{-d}$, (**PolyakSGM**) with $x_0 = \mathbf{0}$, $\eta \leq \frac{1}{\kappa}$, where $\kappa := \frac{\mathsf{L}}{\mu}$, and $f_\star = 0$ produces iterates $\{x_k\}$ satisfying*

$$\|x_{k+1} - x_\star\|^2 \leq \begin{cases} (1 - \frac{\eta\gamma_0}{4\kappa^2})\|x_k - x_\star\|^2, & k < K_0, \\ (1 - \eta\kappa^{-2})\|x_k - x_\star\|^2, & k \geq K_0, \end{cases} \tag{11}$$

where $K_0 = \left\lceil \frac{\kappa^2 \log(4)}{\eta\gamma_0} \right\rceil$, $\gamma_0 = 1/30\pi\|x_\star\|$, $\mathsf{L} \leq 1 + \sqrt{\frac{2d}{m}}$ and $\mu \geq \frac{1}{4\sqrt{\pi}(1+9\pi\|x_\star\|^2)}$.

3. Experiments

We perform a set of synthetic experiments using Gaussian designs. We first plot the convergence behavior of the following 4 methods for different signal scales (see Figure 1):

- (**PolyakSGM**) with standard Polyak step ($\eta = 1$);
- (**PolyakSGM**) with step-size prescribed by Theorem 2.1;
- Gradient descent (GD) with $\eta \in \{2^{-j} \mid -3 \leq j \leq 3\}$ achieving the lowest estimation error;
- Gradient descent (GD) with $\eta = \exp(-5\,\|x_\star\|)$, as prescribed by Theorem 1.1.

We find that (**PolyakSGM**) consistently outperforms GD, while the standard Polyak step ($\eta = 1$) performs best; this suggests that the step-size restriction in Theorem 2.1 might be unnecessary.

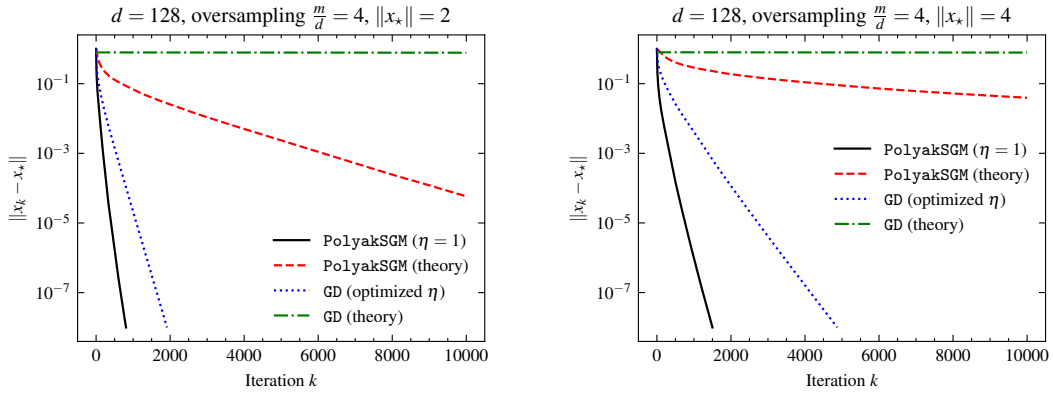

Figure 1: Convergence of (**PolyakSGM**) compared to gradient descent.

We also study the sample efficiency of (**PolyakSGM**) compared to gradient descent. To do so, we generate and solve synthetic problem instances for varying signal scales and oversampling factors. We declare a solve successful if an estimate $\widehat{x}$ satisfying $\|\widehat{x} - x_\star\| \leq 10^{-5}$ is found within 10^4 iterations and calculate the empirical recovery probability over 25 runs (see Figure 2). We observe that gradient descent experiences a sharp cut-off in recovery probability, while (**PolyakSGM**) can recover higher-energy signals from substantially fewer measurements.

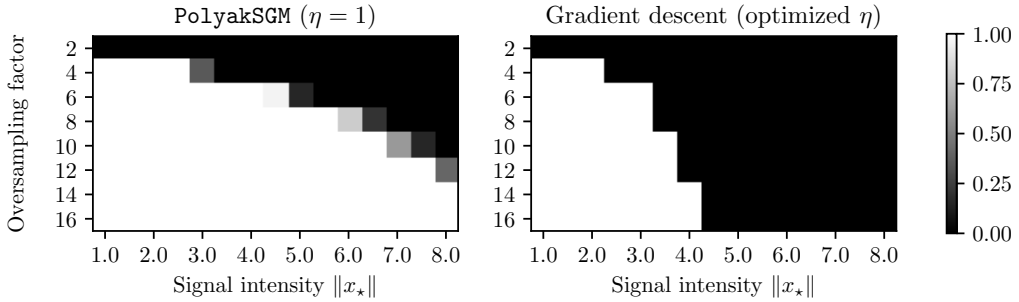

Figure 2: Empirical recovery probability for synthetic instances with $d = 128$. Lighter tiles indicate higher probability of recovery.

References

[1] Rina Foygel Barber and Emil Y Sidky. Convergence for nonconvex ADMM, with applications to CT imaging. *Journal of Machine Learning Research*, 25(38):1–46, 2024.

[2] Marcel Beister, Daniel Kolditz, and Willi A. Kalender. Iterative reconstruction methods in x-ray ct. *Physica Medica*, 28(2):94–108, apr 2012. ISSN 1120-1797. doi: 10.1016/j.ejmp.2012.01.003.

[3] Thorsten M Buzug. Computed tomography. In *Springer handbook of medical technology*, pages 311–342. Springer, 2011.

[4] Center for Nondestructive Evaluation, Iowa State University. Computed tomography. https://www.nde-ed.org/NDETechniques/Radiography/AdvancedTechniques/computedtomography.xhtml, 2024. [Online; accessed on 17-July-2024].

[5] Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-Rank Matrix Recovery with Composite Optimization: Good Conditioning and Rapid Convergence. *Foundations of Computational Mathematics*, 21(6):1505–1593, 2021. ISSN 1615-3383. doi: 10.1007/s10208-020-09490-9.

[6] RK Chhem and DR Brothwell. Paleoradiology. *Imaging mummies and fossils. Berlin: Springer*, 2008.

[7] FH Clarke and Yu S Ledyaev. Mean value inequalities. *Proceedings of the American Mathematical Society*, 122(4):1075–1083, 1994.

[8] Francis H Clarke, Yuri S Ledyaev, Ronald J Stern, and Peter R Wolenski. *Nonsmooth analysis and control theory*, volume 178. Springer Science & Business Media, 2008.

[9] Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

[10] Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20, 2015. ISSN 1083-6489. doi: 10.1214/ejp.v20-3760.

[11] J.A. Fessler and A.O. Hero. Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on Signal Processing*, 42(10):2664–2677, 1994. ISSN 1053-587X. doi: 10.1109/78.324732.

[12] Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of a single neuron with gradient descent. *Advances in Neural Information Processing Systems*, 33:5417–5428, 2020.

[13] Sara Fridovich-Keil, Fabrizio Valdivia, Gordon Wetzstein, Benjamin Recht, and Mahdi Soltanolkotabi. Gradient descent provably solves nonlinear tomographic reconstruction, 2023.

[14] Lin Fu, Tzu-Cheng Lee, Soo Mee Kim, Adam M. Alessio, Paul E. Kinahan, Zhiqian Chang, Ken Sauer, Mannudeep K. Kalra, and Bruno De Man. Comparison Between Pre-Log and Post-Log Statistical Models in Ultra-Low-Dose CT Reconstruction. *IEEE Transactions on Medical Imaging*, 36(3):707–720, March 2017. ISSN 1558-254X. doi: 10.1109/tmi.2016.2627004.

[15] Lucas L Geyer, U Joseph Schoepf, Felix G Meinel, John W Nance Jr, Gorka Bastarrika, Jonathon A Leipsic, Narinder S Paul, Marco Rengo, Andrea Laghi, and Carlo N De Cecco. State of the art: iterative ct reconstruction techniques. *Radiology*, 276(2):339–357, 2015.

[16] Lars Gjesteby, Bruno De Man, Yannan Jin, Harald Paganetti, Joost Verburg, Drosoula Giantsoudi, and Ge Wang. Metal artifact reduction in ct: Where are we after four decades? *IEEE Access*, 4:5826–5849, 2016. doi: 10.1109/ACCESS.2016.2608621.

[17] Jean-Louis Goffin. On convergence rates of subgradient optimization methods. *Mathematical programming*, 13:329–347, 1977.

[18] Richard Gordon. A tutorial on art (algebraic reconstruction techniques). *IEEE Transactions on Nuclear Science*, 21(3):78–93, 1974.

[19] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20(3):451–632, 2020.

[20] F. Natterer. *The Mathematics of Computerized Tomography*. Society for Industrial and Applied Mathematics, 2001. doi: 10.1137/1.9780898719284.

[21] Xiaochuan Pan, Emil Y Sidky, and Michael Vannier. Why do commercial ct scanners still employ traditional, filtered back-projection for image reconstruction? *Inverse Problems*, 25 (12):123009, dec 2009. doi: 10.1088/0266-5611/25/12/123009.

[22] B. T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969. ISSN 0041-5553. doi: 10.1016/0041-5553(69)90061-5.

[23] Johann Radon. On the determination of functions from their integrals along certain manifolds. *Mathematisch-Physische Klasse*, 69:262–277, 1917.

[24] Emil Y Sidky and Xiaochuan Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine and Biology*, 53(17):4777–4807, August 2008. ISSN 1361-6560. doi: 10.1088/0031-9155/53/17/021.

[25] Mahdi Soltanolkotabi. Learning relus via gradient descent. *Advances in neural information processing systems*, 30, 2017.

[26] Mahdi Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *IEEE Transactions on Information Theory*, 65(4):2374–2400, 2019.

[27] Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60 of *A Series of Modern Surveys in Mathematics*. Springer Science & Business Media, 2014. doi: 10.1007/978-3-642-54075-2.

[28] TSA. Computed tomography – Transportation Security Administration. https://www.tsa.gov/computed-tomography, 2024. [Online; accessed on 17-July-2024].

[29] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[30] Martin J. Willemink, Pim A. de Jong, Tim Leiner, Linda M. de Heer, Rutger A. J. Nievelstein, Ricardo P. J. Budde, and Arnold M. R. Schilham. Iterative reconstruction techniques for computed tomography - Part 1: Technical principles. *European Radiology*, 23(6):1623–1631, January 2013. ISSN 1432-1084. doi: 10.1007/s00330-012-2765-y.

Appendix A. Regularity properties of the loss function

A.1. Lipschitz continuity

We first show f is Lipschitz continuous with modulus depending entirely on A.

Proposition A.1 (Lipschitz continuity) *The loss function is Lipschitz continuous:*

$$|f(x) - f(\bar{x})| \leq \left(\frac{1}{m} \sup_{v \in \mathbb{S}^{d-1}} \|Av\|_1 \right) \cdot \|x - \bar{x}\|, \quad \textit{for all } x, \bar{x} \in \mathbb{R}^d. \tag{12}$$

In particular, for Gaussian designs, the following holds with probability at least $1 - 2e^{-cd}$*:*

$$|f(x) - f(\bar{x})| \leq 1 + 2\sqrt{\frac{d}{m}}, \quad \textit{for all } x, \bar{x} \in \mathbb{R}^d.$$

Proof Recall that the absolute value function and $x \mapsto \exp(-x_+)$ are 1-Lipschitz. Therefore,

$$\begin{aligned}
|f(x) - f(\bar{x})| &= \left| \frac{1}{m} \sum_{i=1}^m |y_i - h_i(x)| - |y_i - h_i(\bar{x})| \right| \\
&\leq \frac{1}{m} \sum_{i=1}^m |h_i(x) - h_i(\bar{x})| \\
&= \frac{1}{m} \sum_{i=1}^m \left| e^{-\langle a_i, x \rangle_+} - e^{-\langle a_i, \bar{x} \rangle_+} \right| \\
&\leq \frac{1}{m} \sum_{i=1}^m |\langle a_i, x - \bar{x} \rangle| \\
&\leq \|x - \bar{x}\| \cdot \sup_{u \in \mathbb{S}^{d-1}} \frac{1}{m} \|Au\|_1 ,
\end{aligned}$$

as expected. For Gaussian designs, norm equivalence and [29, Theorem 7.3.3] imply

$$\sup_{u \in \mathbb{S}^{d-1}} \frac{1}{m} \|Au\|_1 \leq \frac{1}{\sqrt{m}} \|A\|_{\mathrm{op}} \leq 1 + 2\sqrt{\frac{d}{m}}$$

with probability at least $1 - 2e^{-cd}$; this completes the proof. ∎

Henceforth, we write $\mathsf{L} := \frac{1}{m} \sup_{v \in \mathbb{S}^{d-1}} \|Av\|_1$ for the Lipschitz modulus. This immediately implies the following upper bound on subgradient norms:

$$\sup_{x \in \mathbb{R}^d} \max_{v \in \partial f(x)} \|v\| \leq \mathsf{L}. \tag{13}$$

A.2. Sharp growth

We continue by establishing that the loss function f grows *sharply* away from its minimizer:

$$f(x) - f(x_\star) \geq \mu \left\| x - x_\star \right\|, \quad \text{for all } x \text{ "near" } x_\star,$$

where μ is given by the following expression:

$$\mu := \frac{1}{4\sqrt{\pi}(1 + 9\pi \left\| x_\star \right\|^2)}. \tag{14}$$

For technical reasons, we prove this claim for all $x \in \mathcal{B}(\mathbf{0}; 2\left\| x_\star \right\|)$, since all iterates of Equation (**PolyakSGM**) initialized at $\mathbf{0}$ remain within that ball. Key to establishing the above claim is the (**Aiming**) inequality, which implies sharpness in a neighborhood of $x_\star$. We now turn to the proof of the aiming inequality.

A.2.1. PROOF OF THE AIMING INEQUALITY

To establish (**Aiming**), we first derive a convenient expression for the inner product.

Lemma A.1 (Subdifferential inner product) *For any point $x \in \mathbb{R}^d$, we have that*

$$\begin{aligned}\langle \partial f(x), x - x_\star \rangle = \frac{1}{m}\sum_{i=1}^m e^{-\langle a_i, x\rangle} \left|\langle a_i, x - x_\star\rangle\right| \mathbb{1}\{\langle a_i, x\rangle > 0\} \\ + \frac{1}{m}\sum_{i=1}^m \mathbb{1}\{\langle a_i, x\rangle = 0\} \left(\langle a_i, x_\star\rangle_+ + \mathbf{sign}(0)\, \langle a_i, x_\star\rangle_-\right).\end{aligned} \tag{15}$$

Proof Recall that $\mathbf{sign}(x) = 1$ for $x > 0$, -1 for $x < 0$, and $\mathbf{sign}(0) = [-1, 1]$. In turn,

$$\begin{aligned}&\mathbf{sign}(y_i - h_i(x)) \\ &= \mathbf{sign}(e^{-\langle a_i, x\rangle_+} - e^{-\langle a_i, x_\star\rangle_+}) \\ &= \mathbf{sign}(e^{-\langle a_i, x\rangle_+} - e^{-\langle a_i, x_\star\rangle_+}) \cdot (\mathbb{1}\{\langle a_i, x - x_\star\rangle \geq 0\} + \mathbb{1}\{\langle a_i, x - x_\star\rangle < 0\})\end{aligned} \tag{16}$$

Note that it suffices to consider $\mathbb{1}\{\langle a_i, x - x_\star\rangle > 0\}$ in the first term of (16), since

$$\frac{1}{m}\sum_{i=1}^m \mathbf{sign}(y_i - h_i(x)) e^{-\langle a_i, x\rangle} \langle a_i, x - x_\star\rangle \mathbb{1}\{\langle a_i, x\rangle \geq 0\}\mathbb{1}\{\langle a_i, x - x_\star\rangle = 0\} = 0.$$

We now proceed on a case-by-case basis.

Case 1: $\langle a_i, x - x_\star\rangle < 0$**.** Since all nonzero terms have $\langle a_i, x\rangle \geq 0$, this means

$$0 \leq \langle a_i, x\rangle < \langle a_i, x_\star\rangle \implies \langle a_i, x_\star\rangle_+ > \langle a_i, x\rangle_+ \implies \mathbf{sign}(e^{-\langle a_i, x\rangle_+} - e^{-\langle a_i, x_\star\rangle_+}) = 1,$$

via monotonicity of the exponential. This yields the partial sum

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^m \mathbf{sign}(y_i - h_i(x))(-e^{-\langle a_i, x\rangle}) \langle a_i, x - x_\star\rangle \mathbb{1}\{\langle a_i, x\rangle \geq 0\}\mathbb{1}\{\langle a_i, x - x_\star\rangle < 0\} \\ &= \frac{1}{m}\sum_{i=1}^m e^{-\langle a_i, x\rangle} \langle a_i, x_\star - x\rangle \mathbb{1}\{\langle a_i, x\rangle \geq 0\}\mathbb{1}\{\langle a_i, x_\star - x\rangle > 0\} \\ &= \frac{1}{m}\sum_{i=1}^m e^{-\langle a_i, x\rangle} \langle a_i, x_\star - x\rangle_+ \mathbb{1}\{\langle a_i, x\rangle \geq 0\}.\end{aligned} \tag{17}$$

Case 2(i): $\langle a_i, x - x_\star \rangle > 0$ **and** $\langle a_i, x \rangle > 0$**.** Note that we have the following possibilities:

- If $\langle a_i, x_\star \rangle < 0$, then $\langle a_i, x_\star \rangle_+ = 0$. Therefore, $\langle a_i, x \rangle > 0$ clearly implies

$$\langle a_i, x \rangle_+ > \langle a_i, x_\star \rangle_+ , \quad \text{and thus} \quad \mathbf{sign}(e^{-\langle a_i, x \rangle_+} - e^{-\langle a_i, x_\star \rangle_+}) = -1.$$

- If $\langle a_i, x_\star \rangle \geq 0$, then $\langle a_i, x_\star \rangle_+ = \langle a_i, x_\star \rangle$; therefore,

$$\langle a_i, x \rangle_+ = \langle a_i, x \rangle > \langle a_i, x_\star \rangle = \langle a_i, x_\star \rangle_+ \implies \mathbf{sign}(e^{-\langle a_i, x \rangle_+} - e^{-\langle a_i, x_\star \rangle_+}) = -1.$$

In either instance, we obtain the partial sum

$$\begin{aligned}
&\frac{1}{m} \sum_{i=1}^m \mathbf{sign}(y_i - h_i(x))(-e^{-\langle a_i, x \rangle}) \, \langle a_i, x - x_\star \rangle \, \mathbb{1}\{\langle a_i, x \rangle > 0\} \mathbb{1}\{\langle a_i, x - x_\star \rangle > 0\} \\
&= \frac{1}{m} \sum_{i=1}^m e^{-\langle a_i, x \rangle} \, \langle a_i, x - x_\star \rangle \, \mathbb{1}\{\langle a_i, x \rangle > 0\} \mathbb{1}\{\langle a_i, x - x_\star \rangle > 0\} \\
&= \frac{1}{m} \sum_{i=1}^m e^{-\langle a_i, x \rangle} \, \langle a_i, x - x_\star \rangle_+ \, \mathbb{1}\{\langle a_i, x \rangle > 0\}. \qquad (18)
\end{aligned}$$

Case 2(ii): $\langle a_i, x - x_\star \rangle > 0$ **and** $\langle a_i, x \rangle = 0$**.** In this case, we have

$$0 = \langle a_i, x \rangle > \langle a_i, x_\star \rangle \implies \langle a_i, x \rangle_+ = \langle a_i, x_\star \rangle_+ = 0.$$

As a result, $\mathbf{sign}(y_i - h_i(x))$ can be any value between $[-1, 1]$. We lower bound

$$\begin{aligned}
&\frac{1}{m} \sum_{i=1}^m \mathbf{sign}(y_i - h_i(x))(-e^{-\langle a_i, x \rangle}) \, \langle a_i, x - x_\star \rangle \, \mathbb{1}\{\langle a_i, x \rangle = 0\} \mathbb{1}\{\langle a_i, x - x_\star \rangle > 0\} \\
&= \frac{1}{m} \sum_{i=1}^m \mathbf{sign}(0) \, \langle a_i, x_\star \rangle \, \mathbb{1}\{\langle a_i, x \rangle = 0\} \mathbb{1}\{\langle a_i, x - x_\star \rangle > 0\} \\
&= \frac{1}{m} \sum_{i=1}^m \mathbf{sign}(0) \cdot \langle a_i, x_\star \rangle_- \, \mathbb{1}\{\langle a_i, x \rangle = 0\}, \qquad (19)
\end{aligned}$$

using $\langle a_i, x_\star \rangle \, \mathbb{1}\{\langle a_i, x_\star \rangle < 0\} = \langle a_i, x_\star \rangle_-$ in (19).

These are all the possible cases to consider. Combining Equations (17) to (19) yields

$$\begin{aligned}
\langle \partial f(x), x - x_\star \rangle &= \frac{1}{m} \sum_{i=1}^m \mathbf{sign}(y_i - h_i(x)) \cdot (-e^{-\langle a_i, x \rangle}) \cdot \langle a_i, x - x_\star \rangle \, \mathbb{1}\{\langle a_i, x \rangle \geq 0\} \\
&= \frac{1}{m} \sum_{i=1}^m e^{-\langle a_i, x \rangle} \, \langle a_i, x - x_\star \rangle \, \mathbb{1}\{\langle a_i, x \rangle > 0\} \mathbb{1}\{\langle a_i, x - x_\star \rangle > 0\} \\
&\quad - \frac{1}{m} \sum_{i=1}^m e^{-\langle a_i, x \rangle} \, \langle a_i, x - x_\star \rangle \, \mathbb{1}\{\langle a_i, x \rangle \geq 0\} \mathbb{1}\{\langle a_i, x - x_\star \rangle < 0\} \\
&\quad + \frac{1}{m} \sum_{i=1}^m \mathbf{sign}(0) \, \langle a_i, x_\star \rangle_- \, \mathbb{1}\{\langle a_i, x \rangle = 0\}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{m}\sum_{i=1}^{m} e^{-\langle a_i, x\rangle} \langle a_i, x - x_\star\rangle \mathbb{1}\{\langle a_i, x\rangle > 0\}\mathbb{1}\{\langle a_i, x - x_\star\rangle \geq 0\} \\
&\quad + \frac{1}{m}\sum_{i=1}^{m} e^{-\langle a_i, x\rangle} \langle a_i, x_\star - x\rangle \mathbb{1}\{\langle a_i, x\rangle \geq 0\}\mathbb{1}\{\langle a_i, x_\star - x\rangle > 0\} \\
&\quad + \frac{1}{m}\sum_{i=1}^{m} \mathbf{sign}(0) \langle a_i, x_\star\rangle_- \mathbb{1}\{\langle a_i, x\rangle = 0\} \\
&= \frac{1}{m}\sum_{i=1}^{m} e^{-\langle a_i, x\rangle} \langle a_i, x - x_\star\rangle_+ \mathbb{1}\{\langle a_i, x\rangle > 0\} \\
&\quad + \frac{1}{m}\sum_{i=1}^{m} e^{-\langle a_i, x\rangle} \langle a_i, x_\star - x\rangle_+ \mathbb{1}\{\langle a_i, x\rangle \geq 0\} \\
&\quad + \frac{1}{m}\sum_{i=1}^{m} \mathbf{sign}(0) \langle a_i, x_\star\rangle_- \mathbb{1}\{\langle a_i, x\rangle = 0\} \\
&= \frac{1}{m}\sum_{i=1}^{m} e^{-\langle a_i, x\rangle} \mathbb{1}\{\langle a_i, x\rangle > 0\} \left(\langle a_i, x - x_\star\rangle_+ + \langle a_i, x_\star - x\rangle_+\right) \\
&\quad + \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{\langle a_i, x\rangle = 0\} \left(\langle a_i, x_\star\rangle_+ + \mathbf{sign}(0) \langle a_i, x_\star\rangle_-\right) \\
&= \frac{1}{m}\sum_{i=1}^{m} e^{-\langle a_i, x\rangle} \mathbb{1}\{\langle a_i, x\rangle > 0\} \cdot |\langle a_i, x - x_\star\rangle| \\
&\quad + \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{\langle a_i, x\rangle = 0\}(\langle a_i, x_\star\rangle_+ + \mathbf{sign}(0) \langle a_i, x_\star\rangle_-),
\end{aligned}
$$

where the last equality follows from the identity $[x]_+ + [-x]_+ = |x|$. ■

With the help of Lemma A.1, we can establish the desired inequality.

Proposition A.2 (Aiming) *If $a_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, the following holds with probability $1 - 2e^{-cd}$:*

$$
\min_{\xi \in \partial f(x)} \langle \xi, x - x_\star\rangle \geq \frac{1}{4\sqrt{\pi}(1 + 9\|x_\star\|^2)} \|x - x_\star\|, \tag{20}
$$

uniformly over all $x \in \mathcal{B}(0; 3\|x_\star\|) \setminus \{0\}$.

Proof We first argue that we can effectively ignore the second term in the expression furnished by Lemma A.1, since it corresponds to a zero-measure event (for any $x \neq 0$). To show this formally, we first note that since $\mathbf{sign}(0) = [-1, 1]$ and $\mathbb{1}\{\langle a_i, x\rangle = 0\} = [0, 1]$ the contribution of the second term is always at least

$$
\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{\langle a_i, x\rangle = 0\} \left(\langle a_i, x_\star\rangle_+ + \mathbf{sign}(0) \langle a_i, x_\star\rangle_-\right) \geq \frac{1}{m}\sum_{i=1}^{m} \mathcal{I}\{\langle a_i, x\rangle = 0\} \langle a_i, x_\star\rangle_-,
$$

writing $\mathcal{I}\{\mathcal{E}\} = 1$ when $\mathcal{E}$ happens and 0 otherwise. We now write

$$\begin{aligned}\mathcal{I}\{\langle a_i, x\rangle = 0\}\langle a_i, x_\star\rangle_- &= \mathcal{I}\{\langle a_i, u\rangle = 0\}\left(\cancelto{0}{\langle a_i, u\rangle}\langle u, x_\star\rangle + \langle P_{u^\perp} a_i, P_{u^\perp} x_\star\rangle\right)_- \\ &\overset{(d)}{=} \mathcal{I}\{\langle a_i, u\rangle = 0\}\cdot\langle \widetilde{a}_i, P_{u^\perp} x_\star\rangle_-, \qquad \widetilde{a}_i \sim \mathcal{N}(0, I_d) \perp\!\!\!\perp a_i,\end{aligned}$$

where we write $u = {}^{x}/_{\|x\|}$, $P_{u^\perp} = (I - uu^\mathsf{T})$, and use the fact that the variables $\langle a_i, u\rangle$ and $P_{u^\perp} a_i$ are uncorrelated (thus independent) Gaussians to replace $P_{u^\perp} a_i$ with an independent copy $P_{u^\perp}\widetilde{a}_i$. From independence, it follows that

$$\begin{aligned}\mathbb{E}\left[\langle a_i, x_\star\rangle_- \mathbb{1}\{\langle a_i, x\rangle = 0\}\right] &\geq \mathbb{E}\left[\mathcal{I}\{\langle a_i, x\rangle = 0\}\cdot\langle a_i, x_\star\rangle_-\right] \\ &= \mathbb{E}\left[\mathcal{I}\{\langle a_i, u\rangle = 0\}\right]\mathbb{E}\left[\langle \widetilde{a}_i, P_{u^\perp} x_\star\rangle_-\right] \\ &= \cancel{\mathbb{P}\left(\langle a_i, u\rangle = 0\right)}\cdot\mathbb{E}\left[\langle \widetilde{a}_i, P_{u^\perp} x_\star\rangle_-\right] \\ &= 0, \end{aligned} \tag{21}$$

where $\mathbb{P}\left(\langle a_i, u\rangle = 0\right)$ since a_i is standard Gaussian and $u \neq 0$. In light of (21), we may ignore the second term in Lemma A.1 in lower-bounding the expectation, since that term has a nonnegative contribution. Continuing from the expression furnished by Lemma (A.1), we obtain

$$\begin{aligned}\langle \partial f(x), x - x_\star\rangle &= \frac{1}{m}\sum_{i=1}^m \exp(-\langle a_i, x\rangle)\,|\langle a_i, x_\star - x\rangle|\,\mathbb{1}\{\langle a_i, x\rangle \geq 0\} \\ &= \frac{\|x - x_\star\|}{m}\sum_{i=1}^m \exp(-\langle a_i, u\rangle\,\|x\|)\,|\langle a_i, v\rangle|\,\mathbb{1}\{\langle a_i, u\rangle \geq 0\} \\ &\geq \frac{\|x - x_\star\|}{m}\sum_{i=1}^m \exp(-3\langle a_i, u\rangle\,\|x_\star\|)\,|\langle a_i, v\rangle|\,\mathbb{1}\{\langle a_i, u\rangle \geq 0\} \\ &\overset{(d)}{=} \frac{\|x - x_\star\|}{m}\sum_{i=1}^m \exp\left(-3\beta_i\,\|x_\star\|\right)\left|\beta_i\langle u, v\rangle + \beta_i^\perp\,\|P_{u^\perp} v\|\right|\mathbb{1}\{\beta_i \geq 0\},\end{aligned}$$

writing $u := \frac{x}{\|x\|}$ and $v := \frac{x_\star - x}{\|x_\star - x\|}$. Here, $\beta_i, \beta_i^\perp \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ arise from the decomposition

$$\begin{aligned}\langle a_i, v\rangle &= \left\langle uu^\mathsf{T} a_i, v\right\rangle + \left\langle (I - uu^\mathsf{T})a_i, v\right\rangle \\ &= \langle a_i, u\rangle\langle u, v\rangle + \left\langle (I - uu^\mathsf{T})a_i, v\right\rangle \\ &\overset{(d)}{=} \langle a_i, u\rangle\langle u, v\rangle + \left\langle (I - uu^\mathsf{T})\widetilde{a}_i, v\right\rangle \qquad (a_i \perp\!\!\!\perp \widetilde{a}_i \sim \mathcal{N}(0, I_d)) \\ &= \langle a_i, u\rangle\langle u, v\rangle + \left\langle \widetilde{a}_i, (I - uu^\mathsf{T})v\right\rangle \\ &\overset{(d)}{=} \beta_i\langle u, v\rangle + \beta_i^\perp\|(I - uu^\mathsf{T})v\|,\end{aligned}$$

writing $\beta_i = \langle a_i, u\rangle \sim \mathcal{N}(0,1)$ and using the identity $\langle \widetilde{a}_i, z\rangle \sim \mathcal{N}(0, \|z\|^2)$; we also recognize $(I - uu^\mathsf{T}) = P_{u^\perp}$. We now consider two cases for the correlation $\langle u, v\rangle$:

Case 1: $\langle u, v\rangle \geq 0$**.** In this case, we may lower bound the sum by the following expression:

$$\langle \partial f(x), x - x_\star\rangle$$

$$\geq \frac{\|x - x_\star\|}{m} \sum_{i=1}^{m} \exp(-3\beta_i \|x_\star\|) \left|\beta_i \langle u, v\rangle + \beta_i^\perp \|P_{u^\perp} v\|\right| \mathbb{1}\{\beta_i \geq 0, \beta_i^\perp \geq 0\}$$
$$= \frac{\|x - x_\star\|}{m} \sum_{i=1}^{m} \exp(-3\beta_i \|x_\star\|) \left(\beta_i |\langle u, v\rangle| + \beta_i^\perp \|P_{u^\perp} v\|\right) \mathbb{1}\{\beta_i \geq 0, \beta_i^\perp \geq 0\}. \tag{22}$$

Case 2: $\langle u, v\rangle < 0$**.** In this case, we may lower bound the sum by the following expression:

$$\begin{aligned}
&\langle \partial f(x), x - x_\star\rangle \\
&\geq \frac{\|x - x_\star\|}{m} \sum_{i=1}^{m} \exp(-3\beta_i \|x_\star\|) \left|\beta_i \langle u, v\rangle + \beta_i^\perp \|P_{u^\perp} v\|\right| \mathbb{1}\{\beta_i \geq 0, \beta_i^\perp \leq 0\} \\
&= \frac{\|x - x_\star\|}{m} \sum_{i=1}^{m} \exp(-3\beta_i \|x_\star\|) \left(\beta_i(-\langle u, v\rangle) - \beta_i^\perp \|P_{u^\perp} v\|\right) \mathbb{1}\{\beta_i \geq 0, \beta_i^\perp \leq 0\} \\
&\overset{(d)}{=} \frac{\|x - x_\star\|}{m} \sum_{i=1}^{m} \exp(-3\beta_i \|x_\star\|) \left(\beta_i |\langle u, v\rangle| + \widetilde{\beta}_i^\perp \|P_{u^\perp} v\|\right) \mathbb{1}\{\beta_i, \widetilde{\beta}_i^\perp \geq 0\},
\end{aligned} \tag{23}$$

where $\widetilde{\beta}_i \sim \mathcal{N}(0, 1) \perp\!\!\!\perp \beta_i$, using the fact that the expression inside the absolute value is nonpositive. Since the sum in (23) is distributionally identical to (22), it suffices to study

$$(\natural) := \frac{1}{m} \sum_{i=1}^{m} \exp(-3\beta_i \|x_\star\|) \left(\beta_i |\langle u, v\rangle| + \beta_i^\perp \|P_{u^\perp} v\|\right) \mathbb{1}\{\beta_i, \beta_i^\perp \geq 0\}. \tag{24}$$

Taking expectations with respect to β and $\beta^\perp$ and writing $\gamma := 3\|x_\star\|$ for brevity, we obtain

$$\begin{aligned}
&\mathbb{E}_{\beta,\beta^\perp}\left[\exp(-\gamma\beta)\left(\beta|\langle u, v\rangle| + \beta^\perp \|P_{u^\perp} v\|\right) 1\{\beta, \beta^\perp \geq 0\}\right] \\
&= |\langle u, v\rangle| \cdot \mathbb{E}_\beta\left[\exp(-\gamma\beta)\beta_+\right] \cdot \mathbb{P}\left(\beta^\perp \geq 0\right) + \|P_{u^\perp} v\| \, \mathbb{E}_{\beta^\perp}[\beta_+^\perp] \cdot \mathbb{E}_\beta\left[\exp(-\gamma\beta)\mathbb{1}\{\beta \geq 0\}\right] \\
&= \frac{|\langle u, v\rangle|}{4}\left(\sqrt{\frac{2}{\pi}} - \gamma \exp\left(\frac{\gamma^2}{2}\right) \operatorname{erfc}\left(\frac{\gamma}{\sqrt{2}}\right)\right) + \frac{\|P_{u^\perp} v\|}{4}\sqrt{\frac{2}{\pi}} \exp\left(\frac{\gamma^2}{2}\right) \operatorname{erfc}\left(\frac{\gamma}{\sqrt{2}}\right) \\
&\geq \frac{|\langle u, v\rangle|}{4(1 + \pi\gamma^2)} \cdot \sqrt{\frac{2}{\pi}} + \frac{\|P_{u^\perp} v\|}{4}\sqrt{\frac{2}{\pi}} \exp\left(\frac{\gamma^2}{2}\right) \operatorname{erfc}\left(\frac{\gamma}{\sqrt{2}}\right) \\
&= \frac{|\langle u, v\rangle|}{4(1 + \pi\gamma^2)} \cdot \sqrt{\frac{2}{\pi}} + \frac{\|P_{u^\perp} v\|}{4\gamma}\sqrt{\frac{2}{\pi}} \cdot \gamma \exp\left(\frac{\gamma^2}{2}\right) \operatorname{erfc}\left(\frac{\gamma}{\sqrt{2}}\right) \\
&\geq \sqrt{\frac{1}{8\pi}} \cdot \min\{\frac{1}{1 + \pi\gamma^2}, \frac{1}{2\gamma}\} \left(|\langle u, v\rangle| + \|P_{u^\perp} v\|\right) \\
&\geq \sqrt{\frac{1}{8\pi}} \cdot \frac{1}{1 + \pi\gamma^2} \|P_u v + P_{u^\perp} v\| \\
&= \sqrt{\frac{1}{8\pi}} \frac{1}{1 + \pi\gamma^2},
\end{aligned}$$

using Lemma E.1 in the second equality, Lemma E.4 in the first inequality, Lemma E.3 and the bound $\gamma \geq 1$ in the second inequality, the fact that $|\langle u, v\rangle| = \|P_u v\|$ and the triangle inequality in

the last inequality, and $v \in \mathbb{S}^{d-1}$ in the last step. As a result, we have

$$\mathbb{E}[\langle \partial f(x), x - x_\star \rangle] \geq \frac{\|x - x_\star\|}{\sqrt{8\pi}\left(1 + 9\pi \|x_\star\|^2\right)}, \quad \text{for any } x \in \mathcal{B}(\mathbf{0}; 3\|x_\star\|). \tag{25}$$

Given (25), we finish off the proof with a concentration argument.

High-probability version. Our proof relies on the generic chaining technique [27] using the refinements of [10]. Setting the stage, consider a random process indexed by a set $\mathcal{T}$, $(Z_t)_{t\in\mathcal{T}}$. We say that (Z_t) has *mixed tails* with respect to metrics (d_1, d_2) if

$$\mathbb{P}\left(|Z_t - Z_s| \geq u d_1(s,t) + \sqrt{u} d_2(s,t)\right) \leq e^{-u}. \tag{26}$$

Under (26), [10, Theorem 3.5] shows that for any $t_0 \in \mathcal{T}$, it holds that

$$\mathbb{P}\left(\sup_{t\in\mathcal{T}} |Z_t - Z_{t_0}| \geq C(\gamma_1(\mathcal{T}, d_1) + \gamma_2(\mathcal{T}, d_2)) + u\,\mathbf{diam}_{d_1}(\mathcal{T}) + \sqrt{u}\,\mathbf{diam}_{d_2}(\mathcal{T})\right) \leq e^{-u}, \tag{27}$$

where $\mathbf{diam}_d(\mathcal{T}) := \sup_{s,t\in\mathcal{T}} d(s,t)$ and $\gamma_\alpha(\mathcal{T}, d)$ is the Talagrand's γ-*functional*:

$$\gamma_\alpha(\mathcal{T}, d) = \inf_{(\mathcal{T}_k)} \sup_{t\in\mathcal{T}} \sum_{k=0}^{\infty} 2^{k/\alpha} d(t, \mathcal{T}_k), \qquad \mathcal{T}_k \in \{\mathcal{H} \subset \mathcal{T} \mid |\mathcal{H}| \leq 2^{2^k}\}.$$

The first step in our proof is to verify condition (26) for the random process at hand.

Claim 1 (Mixed tail condition) *For any $(u, v) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ and $i \in [m]$, let*

$$Z_i(u,v) := \exp\left(-3\langle a_i, u\rangle \|x_\star\|\right) |\langle a_i, v\rangle| \, \mathbb{1}\{\langle a_i, u\rangle > 0\} + \mathcal{I}\{\langle a_i, u\rangle = 0\} \langle a_i, x_\star\rangle_- . \tag{28}$$

Then the process $Z(u,v) := \frac{1}{m}\sum_{i=1}^m Z_i(u,v) - \mathbb{E}[Z_i(u,v)]$ satisfies (26) with

$$d_1 = \frac{(1 + 3\|x_\star\|) d_{\mathsf{euc}}}{m}, \quad d_2 = \frac{(1 + 3\|x_\star\|) d_{\mathsf{euc}}}{\sqrt{m}},$$

where d_{euc} denotes the Euclidean metric on $\mathbb{R}^d \times \mathbb{R}^d$.

With Claim 1 at hand, we can invoke (27) for $\mathcal{T} := (\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}) \cup \{\mathbf{0}\}$:

$$\mathbb{P}\left(\sup_{u,v} |Z(u,v)| \geq C(\gamma_1(\mathcal{T}, d_1) + \gamma_2(\mathcal{T}, d_2)) + t\,\mathbf{diam}_{d_1}(\mathcal{T}) + \sqrt{t}\,\mathbf{diam}_{d_2}(\mathcal{T})\right) \leq 2e^{-t}. \tag{29}$$

To simplify (29), we bound the γ-functionals using Dudley's entropy integral method.

Claim 2 (Dudley bounds I) *Let $\mathcal{T} := (\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}) \cup \{\mathbf{0}\}$. We have*

$$\mathbf{diam}_{d_1}(\mathcal{T}) \lesssim \frac{(1 + 3\|x_\star\|)}{m}, \quad \gamma_1(\mathcal{T}, d_1) \lesssim \frac{(1 + 3\|x_\star\|) d}{m}. \tag{30}$$

Claim 3 (Dudley bounds II) *Let $\mathcal{T} := (\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}) \cup \{\mathbf{0}\}$. We have*

$$\mathbf{diam}_{d_2}(\mathcal{T}) \lesssim \frac{(1 + 3\|x_\star\|)}{\sqrt{m}}, \quad \gamma_2(\mathcal{T}, d_2) \lesssim (1 + 3\|x_\star\|)\sqrt{\frac{d}{m}}. \tag{31}$$

Combining Equation (29) and Claims 2 and 3, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\sup_{u,v}|Z(u,v)| \geq C(1+3\|x_\star\|)\left(\sqrt{\frac{d}{m}}+\frac{d}{m}\right)\right)\\
&\leq \mathbb{P}\left(\sup_{u,v}|Z(u,v)| \geq C\left(\gamma_1(\mathcal{T},d_1)+\gamma_2(\mathcal{T},d_2)\right)+d\cdot\mathbf{diam}_{d_1}(\mathcal{T})+\sqrt{d}\cdot\mathbf{diam}_{d_2}(\mathcal{T})\right)\\
&\leq 2e^{-d}.
\end{aligned}$$

By definition of $Z(u,v)$ and (25), it follows that

$$\frac{1}{m}\sum_{i=1}^{m} e^{-3\langle a_i,u\rangle\|x_\star\|}\,|\langle a_i,v\rangle|\,\mathbb{1}\{\langle a_i,u\rangle\geq 0\} \geq \frac{1}{\sqrt{8\pi}\left(1+9\pi\|x_\star\|^2\right)} - C\left(\sqrt{\frac{d}{m}}+\frac{d}{m}\right),$$

uniformly over $u,v\in\mathbb{S}^{d-1}$, with probability $1-2e^{-d}$. Therefore, we conclude that

$$\min_{v\in\partial f(x)}\langle v, x-x_\star\rangle \geq \frac{\|x-x_\star\|}{4\sqrt{\pi}\left(1+9\pi\|x_\star\|^2\right)},\quad \forall x\in\mathcal{B}(\mathbf{0};3\|x_\star\|)\setminus\{0\},$$

with probability at least $1-2e^{-d}$, as long as $m\gtrsim d\cdot(1+9\pi\|x_\star\|^2)^2\gtrsim d\|x_\star\|^4$. Finally, we place the proofs of Claims 1 to 3 in Appendix D. ■

Appendix B. Convergence analysis

In this section, we analyze the convergence of (**PolyakSGM**). First, we denote

$$\overline{\rho}:=\eta\left(\frac{\mu}{\mathsf{L}}\right)^2\frac{\gamma_0}{4},\quad \rho:=\eta\left(\frac{\mu}{\mathsf{L}}\right)^2,\quad \gamma_0=\frac{1}{30\pi\|x_\star\|},\quad \text{and}\quad T_0:=\left\lceil\frac{\log(4)}{\overline{\rho}}\right\rceil \tag{32}$$

for brevity. A quick discussion about their roles in the convergence analysis is in order:

- $\overline{\rho}$ is the contraction factor achieved while $\|x_t-x_\star\|\geq\frac{1}{2}\|x_\star\|$.
- $\rho>\overline{\rho}$ is the contraction factor achieved after $\|x_t-x_\star\|<\frac{1}{2}\|x_\star\|$.
- T_0 is an upper bound on the number of iterations elapsed until $\|x_t-x_\star\|<\frac{1}{2}\|x_\star\|$.

As the above quantities suggest, our convergence analysis outlines a "slow" and "fast" phase of convergence (albeit both following a geometric rate). Crucially, Equation (**PolyakSGM**) **does not** employ a different stepsize for the two phases: the distinction is only for theoretical purposes and – as demonstrated in our numerical experiments – does not affect the practical behavior of the method. For that reason, we believe it is just an artifact of our analysis.

We now turn to the analysis of the algorithm. We define the following events:

$$\mathcal{A}_{\mathsf{slow}}(t):=\left\{\|x_{t+1}-x_\star\|^2\leq(1-\overline{\rho})\|x_t-x_\star\|^2\right\}, \tag{33a}$$
$$\mathcal{A}_{\mathsf{fast}}(t):=\left\{\|x_{t+1}-x_\star\|^2\leq(1-\rho)\|x_t-x_\star\|^2\right\}, \tag{33b}$$
$$\mathcal{B}_{\mathsf{slow}}(t):=\left\{\frac{\gamma_0}{2}\leq\|x_t\|\leq 2\|x_\star\|\right\},\quad \mathcal{B}_{\mathsf{fast}}(t):=\left\{\frac{1}{2}\|x_\star\|\leq\|x_t\|\leq 2\|x_\star\|\right\}. \tag{33c}$$

We also recall that $x_0 = \mathbf{0}$ and $f_\star = 0$ throughout. Our analysis is inductive: initially,

$$\|x_1 - x_\star\|^2 \le (1 - \gamma_0) \|x_\star\|^2, \quad \text{where} \quad \gamma_0 := \frac{1}{30\pi \|x_\star\|}. \tag{34}$$

The following Lemma contains a formal proof of (34).

Lemma B.1 (Properties at initialization) *Suppose* $a_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ *and let*

$$v_0 = -\frac{1}{m} \sum_{i=1}^m a_i \mathbb{1}\{\langle a_i, x_\star \rangle > 0\} \in \partial f(x_0).$$

Then, if $m \gtrsim d \|x_\star\|^2$ *and* $\eta \le \frac{1}{3\sqrt{\pi}}$*, the following holds with probability* $1 - 2e^{-d}$*:*

$$\langle v_0, x_0 - x_\star \rangle \ge \|x_\star\| \left(\sqrt{\frac{1}{2\pi}} - \sqrt{\frac{2d}{m}} \right), \tag{35a}$$

$$\|x_1 - x_\star\|^2 \le \left(1 - \frac{\eta}{10\sqrt{\pi} \|x_\star\|} \right) \|x_\star\|^2. \tag{35b}$$

Proof The inclusion $v_0 \in \partial f(x_0)$ can be verified by the chain rule. For (35a), we write

$$\begin{aligned} \langle v_0, x_0 - x_\star \rangle &= \frac{1}{m} \sum_{i=1}^m \langle a_i, x_\star \rangle_+ \\ &\overset{(d)}{=} \|x_\star\| \cdot \frac{1}{m} \sum_{i=1}^m (\beta_i)_+, \end{aligned}$$

where $\beta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. For the latter sum, we calculate

$$\mathbb{E}\left[\frac{1}{m} \sum_{i=1}^m (\beta_i)_+ \right] = \sqrt{\frac{1}{2\pi}}.$$

Moreover, the sum is $\frac{1}{\sqrt{m}}$-Lipschitz with respect to the random parameters:

$$\left| \frac{1}{m} \sum_{i=1}^m (\beta_i)_+ - \frac{1}{m} \sum_{i=1}^m (\bar{\beta}_i)_+ \right| \le \frac{1}{m} \sum_{i=1}^m |\beta_i - \bar{\beta}_i| = \frac{1}{m} \|\beta - \bar{\beta}\|_1 \le \frac{1}{\sqrt{m}} \|\beta - \bar{\beta}\|,$$

using the fact that $x \mapsto \max(x, 0)$ is 1-Lipschitz and norm equivalence. Consequently, the Gaussian Lipschitz concentration inequality [29, Theorem 5.2.2] yields

$$\mathbb{P}\left(\left| \langle v_0, x_0 - x_\star \rangle - \frac{\|x_\star\|}{\sqrt{2\pi}} \right| \ge \|x_\star\| \cdot t \right) \le 2 \exp\left(-\frac{mt^2}{2} \right).$$

Setting $t = \sqrt{\frac{2d}{m}}$ above yields (35a). For (35b), we will require an auxiliary result, whose proof is deferred to Appendix D.

Claim 4 *Suppose that* $m \gtrsim d \|x_\star\|^2$ *and* $\|x_\star\| \geq 1$. *Then*

$$\mathbb{P}\left(f(0) \leq \frac{1}{5}\right) \leq \exp\left(-d\right). \tag{36}$$

With Claim 4 at hand, we have the following chain of inequalities:

$$\begin{aligned}
&\|x_1 - x_\star\|^2 \\
&= \|x_\star\|^2 + \frac{\eta^2 f(0)^2}{\|v_0\|^2} + 2\eta \frac{f(0)}{\|v_0\|^2} \langle v_0, x_\star \rangle \\
&\leq \|x_\star\|^2 - \frac{\eta f(0)}{\|v_0\|^2} \left(2\left(\sqrt{\frac{1}{2\pi}} - \sqrt{\frac{2d}{m}}\right)\|x_\star\| - \eta \mathsf{L} \|x_\star\|\right) && \text{((35a) + Prop. A.1)} \\
&\leq \|x_\star\|^2 - \frac{\eta f(0)}{\|v_0\|^2} \frac{1}{\sqrt{\pi}} \|x_\star\| && \left(m \gtrsim d \text{ and } \eta \leq \frac{1}{3\sqrt{\pi}}\right) \\
&\leq \|x_\star\|^2 \left(1 - \frac{\eta}{5\mathsf{L}\sqrt{\pi}\|x_\star\|}\right) && \text{((35b))} \\
&\leq \|x_\star\|^2 \left(1 - \frac{\eta}{10\sqrt{\pi}\|x_\star\|}\right),
\end{aligned}$$

using $\mathsf{L} \leq 2$ for $m \gtrsim d$ in the last inequality. ■

We now turn to a sequence of supporting Lemmas. The first one shows that the norms of the iterates remain bounded while the algorithm is in its "slow" phase.

Lemma B.2 *We have that* $\{\mathcal{A}_{\mathsf{slow}}(j)\}_{j \leq t} \implies \mathcal{B}_{\mathsf{slow}}(t+1)$.

Proof We have the following chain of inequalities:

$$\begin{aligned}
\|x_{t+1} - x_\star\|^2 &\overset{(\mathcal{A}_{\mathsf{slow}}(t))}{\leq} (1 - \bar{\rho}) \|x_t - x_\star\|^2 \\
&\overset{(\mathcal{A}_{\mathsf{slow}}(j))_{1<j<t}}{\leq} (1 - \bar{\rho})^{t-1} \|x_1 - x_\star\|^2 \\
&\overset{(\mathcal{A}_{\mathsf{slow}}(1))}{\leq} (1 - \bar{\rho})^{t-1} (1 - \gamma_0) \|x_\star\|^2.
\end{aligned}$$

Consequently, taking off squares and using the inequality $\sqrt{1-x} \leq 1 - \frac{x}{2}$ yields

$$\|x_{t+1} - x_\star\| \leq \left(1 - \frac{\gamma_0}{2}\right) \|x_\star\|.$$

Invoking Claim 6 with $\delta := \gamma_0/2$ yields the result. ■

The next Lemma shows the algorithm continues making progress (at a rate depending on $\bar{\rho}$) while the iterates remain within the tube $\mathcal{B}(\mathbf{0}; 2\|x_\star\|) \setminus \mathcal{B}(\mathbf{0}; \gamma_0/2)$.

Lemma B.3 *We have that* $\{\mathcal{A}_{\mathsf{slow}}(j)_{j<t}, \mathcal{B}_{\mathsf{slow}}(t)\} \implies \mathcal{A}_{\mathsf{slow}}(t)$ *for any* $\eta \leq \frac{\mu}{\mathsf{L}}$ *and* $t \geq 1$.

Proof Note that $\mathcal{B}_{\mathsf{slow}}(t)$ guarantees $x_t \in \mathcal{B}(0; 2\|x_\star\|) \setminus \{0\}$. By (**Aiming**) and (8)

$$f(x_t) - f_\star \geq \mu \cdot \min\{\|x_t - x_\star\|, \|x_t\|\}, \quad \text{and} \quad \langle v_t, x_t - x_\star \rangle \geq \mu \|x_t - x_\star\|.$$

At the same time, the events $\mathcal{A}_{\mathsf{slow}}(j)_{j<t}$ guarantee that

$$\|x_t - x_\star\|^2 \leq (1-\bar{\rho})^{t-1}(1-\gamma_0)\|x_\star\|^2 \leq (1-\gamma_0)\|x_\star\|^2.$$

Invoking Lemma E.5 with $\delta := \gamma_0$ leads to

$$f(x_t) - f_\star \geq \mu \cdot \min\{\|x_t - x_\star\|, \|x_t\|\} \geq \mu \cdot \frac{\gamma_0}{4}\|x_t - x_\star\|. \tag{37}$$

From Lemma A.1 and Equation (37), it follows that

$$\begin{aligned}
\|x_{t+1} - x_\star\|^2 &\leq \left\| x_t - x_\star - \eta \frac{f(x_t)}{\|v_t\|^2} v_t \right\|^2 \\
&= \|x_t - x_\star\|^2 - \frac{\eta f(x_t)}{\|v_t\|^2}\left(2\langle v_t, x_t - x_\star\rangle - \eta f(x_t)\right) \\
&\leq \|x_t - x_\star\|^2 - \frac{\eta f(x_t)}{\|v_t\|^2}\left(2\mu\|x_t - x_\star\| - \eta \mathsf{L}\|x_t - x_\star\|\right) \\
&\leq \|x_t - x_\star\|^2 - \eta\left(\frac{\mu}{\mathsf{L}}\right)^2 \frac{\gamma_0}{4}\|x_t - x_\star\|^2 \\
&= \|x_t - x_\star\|^2 (1-\bar{\rho}),
\end{aligned}$$

using the range of η in the penultimate inequality. This proves $\mathcal{A}_{\mathsf{slow}}(t)$. ■

The forthcoming Lemmas describe the behavior of the algorithm once iterates are in the ball $\mathcal{B}(x_\star; \frac{1}{2}\|x_\star\|)$. Lemma B.4 shows that the algorithm indeed enters this region after sufficient progress; the remaining Lemmas mirror Lemmas B.2 and B.3.

Lemma B.4 *We have* $\{\mathcal{A}_{\mathsf{slow}}(j)\}_{j<T_0} \implies \mathcal{B}_{\mathsf{fast}}(T_0)$.

Proof Iterating the inequalities furnished by the events $\mathcal{A}_{\mathsf{slow}}(j)$ for $j < T_0$ yields

$$\|x_{T_0} - x_\star\|^2 \leq (1-\bar{\rho})^{T_0}\|x_\star\|^2 \leq \exp(-T_0\bar{\rho})\|x_\star\|^2 \leq \frac{1}{4}\|x_\star\|^2,$$

using the fact that $T_0 = \lceil \log(4)/\bar{\rho} \rceil$. Appealing to Claim 6 with $\delta = 1/2$ implies

$$\|x_t\| = \sup_{u \in \mathbb{S}^{d-1}} \langle x_t, u\rangle \geq \frac{1}{\|x_\star\|}\langle x_t, x_\star\rangle \geq \frac{1}{2}\|x_\star\|,$$

showing the lower bound in $\mathcal{B}_{\mathsf{fast}}(T_0)$. For the upper bound, the reverse triangle inequality yields

$$\|x_{T_0} - x_\star\| < \frac{1}{2}\|x_\star\| \implies \|x_{T_0}\| \leq \left(1 + \frac{1}{2}\right)\|x_\star\|,$$

as expected. This completes the proof of $\mathcal{B}_{\mathsf{fast}}(T_0)$. ■

Lemma B.5 *We have that* $\{\mathcal{A}_{\mathsf{fast}}(j)_{T_0\le j<t}, \mathcal{B}_{\mathsf{fast}}(t)\} \implies \mathcal{A}_{\mathsf{fast}}(t)$ *for any* $\eta \le \frac{\mu}{\mathsf{L}}$ *and* $t \ge T_0$.

Proof We start by noting that $\mathcal{B}_{\mathsf{fast}}(t)$ combined with (8) guarantee that

$$f(x_t) - f_\star \ge \mu \cdot \min\{\|x_t\|, \|x_t - x_\star\|\} \ge \mu \cdot \min\Big\{\frac{1}{2}\|x_\star\|, \|x_t - x_\star\|\Big\}.$$

Concurrently, $\{\mathcal{A}_{\mathsf{fast}}(j)\}_{T_0\le j<t}$ implies $\|x_t - x_\star\| \le \|x_{T_0} - x_\star\| \le \frac{1}{2}\|x_\star\|$. Consequently,

$$f(x_t) - f_\star \ge \mu \|x_t - x_\star\|.$$

The rest follows *mutatis-mutandis* from the proof of Lemma B.3. ■

Lemma B.6 *We have* $\{\mathcal{A}_{\mathsf{fast}}(j)\}_{T_0\le j\le t} \implies \mathcal{B}_{\mathsf{fast}}(t+1)$.

Proof On the events $\mathcal{A}_{\mathsf{fast}}(j)$ for $T_0 \le j \le t$, we have

$$\|x_{t+1} - x_\star\| \le (1-\rho)^{t-T_0}\|x_{T_0} - x_\star\| \le \frac{1}{2}\|x_\star\|.$$

We deduce the lower bound in $\mathcal{B}_{\mathsf{fast}}(t+1)$ from the preceding display and Claim 6; the upper bound follows from the reverse triangle inequality. ■

Appendix C. Clarke subdifferential of the loss

In this section, we provide an explicit calculation of $\partial f(x)$. With the help of the chain rule, we have

$$\partial f(x) = \frac{1}{m}\sum_{i=1}^m \mathbf{sign}(y_i - h_i(x)) \cdot (-e^{-\langle a_i, x\rangle_+}) a_i \mathbb{1}\{\langle a_i, x\rangle \ge 0\}, \tag{38a}$$

$$\text{where} \quad \mathbf{sign}(x) := \begin{cases} 1, & x > 0, \\ -1, & x < 0, \\ [-1,1], & x = 0, \end{cases} \quad \text{and} \quad \mathbb{1}\{x \ge 0\} := \begin{cases} 1, & x > 0 \\ 0, & x < 0, \\ [0,1] & x = 0. \end{cases} \tag{38b}$$

Remark 1 *Our implementation uses* $\mathbf{sign}(0) = 0$ *and* $\mathbb{1}\{x \ge 0\} = 1$ *if* $x = 0$ *in* (38b).

Appendix D. Omitted proofs

Proof of Claim 1. Let (u, v) and $(\bar{u}, \bar{v}) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ and consider the difference $Z_i(u,v) - Z_i(\bar{u},\bar{v})$. Writing $\gamma := 3\|x_\star\|$, we obtain

$$\begin{aligned} \|Z_i(u,v) - Z_i(\bar{u},\bar{v})\|_{\psi_1} \le\ & \big\|(|\langle a_i, v\rangle| - |\langle a_i, \bar{v}\rangle|)e^{-\langle a_i, u\rangle\gamma}\mathbb{1}\{\langle a_i, u\rangle \ge 0\}\big\|_{\psi_1} \\ & + \big\|(e^{-\langle a_i,u\rangle\gamma}\mathbb{1}\{\langle a_i,u\rangle \ge 0\} - e^{-\langle a_i,\bar{u}\rangle\gamma}\mathbb{1}\{\langle a_i,\bar{u}\rangle \ge 0\})\,|\langle a_i,\bar{v}\rangle|\big\|_{\psi_1} \\ & + \|\langle a_i, x_\star\rangle_-\,(\mathcal{I}\{\langle a_i,u\rangle = 0 - \mathcal{I}\{\langle a_i,\bar{u}\rangle = 0\})\|_{\psi_1} \\ \le\ & \||\langle a_i, v-\bar{v}\rangle|\|_{\psi_1} + \gamma\,\||\langle a_i, u-\bar{u}\rangle|\,|\langle a,\bar{v}\rangle|\|_{\psi_1} \end{aligned}$$

$$\leq \|v - \bar{v}\| + \|u - \bar{u}\| \cdot \gamma$$
$$\lesssim (1+\gamma) \left\| \begin{bmatrix} u \\ v \end{bmatrix} - \begin{bmatrix} \bar{u} \\ \bar{v} \end{bmatrix} \right\|,$$

using the fact that the mapping $x \mapsto e^{-\gamma x}\mathbb{1}\{x \geq 0\}$ is γ-Lipschitz on $x \geq 0$, as well as the property $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}$ [29, Lemma 2.7.7]; in particular, using the latter property implies that the last term in the decomposition satisfies

$$\begin{aligned} \|\langle a_i, x_\star \rangle_- \left(\mathcal{I}\{\langle a_i, u\rangle = 0\} - \mathcal{I}\{\langle a_i, \bar{u}\rangle = 0\}\right)\|_{\psi_1} &\lesssim \|\mathcal{I}\{\langle a_i, u\rangle = 0\}\|_{\psi_2} + \|\mathcal{I}\{\langle a_i, \bar{u}\rangle = 0\}\|_{\psi_2} \\ &= 0, \end{aligned}$$

where the last line follows from [29, Proposition 2.5.2] and the identity

$$\begin{aligned} \|\mathcal{I}\{\langle a_i, u\rangle = 0\}\|_{L^p} &= \left(\mathbb{E}\left[\left(\mathcal{I}\{\langle a_i, u\rangle = 0\}\right)^p\right]\right)^{\frac{1}{p}} \\ &= \left(\mathbb{E}\left[\mathcal{I}\{\langle a_i, u\rangle = 0\}\right]\right)^{\frac{1}{p}} \\ &= 0. \end{aligned}$$

From the Bernstein inequality [29, Corollary 2.8.3], it follows that

$$\mathbb{P}\left(|Z(u,v) - Z(\bar{u},\bar{v})| \geq t\right) \leq 2e^{-c\min\left\{\frac{mt^2}{(1+\gamma)^2 d_{\mathsf{euc}}^2((u,v),(\bar{u},\bar{v}))}, \frac{mt}{(1+\gamma)d_{\mathsf{euc}}((u,v),(\bar{u},\bar{v}))}\right\}}, \tag{39}$$
$$\text{where} \quad Z(u,v) := \frac{1}{m}\sum_{i=1}^m Z_i(u,v) - \mathbb{E}[Z_i(u,v)].$$

To argue that $Z(u,v)$ has mixed tails, we let

$$d_1 = \frac{(1+\gamma)d_{\mathsf{euc}}}{m}, \quad d_2 = \frac{(1+\gamma)d_{\mathsf{euc}}}{\sqrt{m}}, \quad s = t \cdot d_1((u,v),(\bar{u},\bar{v})) + \sqrt{t} \cdot d_2((u,v),(\bar{u},\bar{v})).$$

Substituting in (39), and noting $s \geq t d_1((u,v),(\bar{u},\bar{v}))$ and $s^2 \geq t d_2^2((u,v),(\bar{u},\bar{v}))$, we obtain

$$\mathbb{P}\left(|Z(u,v) - Z(\bar{u},\bar{v})| \geq s\right) \leq 2\exp\left(-c\min\left\{\frac{s^2}{d_2^2}, \frac{s}{d_1}\right\}\right) \leq 2e^{-ct}.$$

Relabeling and adjusting constants yields the mixed tail condition. ■

Proof of Claim 2. Using Dudley's entropy integral method yields

$$\begin{aligned} \gamma_1(\mathcal{T}, d_1) &= \frac{1+\gamma}{m}\gamma_1(\mathcal{T}, \|\cdot\|) \\ &\lesssim \frac{1+\gamma}{m}\int_0^\infty \log_+ \mathcal{N}\left(\mathcal{T}, u\right) \mathrm{d}u \\ &= \frac{1+\gamma}{m}\int_0^\infty \log_+ \mathcal{N}(\mathcal{T}, u)\mathrm{d}u \\ &= \frac{1+\gamma}{m}\int_0^1 \log \mathcal{N}(\mathcal{T}, u)\mathrm{d}u \end{aligned}$$

$$\lesssim \frac{(1+\gamma)d}{m}\int_0^1 \log\frac{1}{\varepsilon}\mathrm{d}\varepsilon$$
$$= \frac{(1+3\left\|x_\star\right\|)d}{m},$$

using the fact that the ϵ-covering number of the unit sphere is $(1/\epsilon)^d$, thus the covering number of their Cartesian product is $(1/\epsilon)^{2d}$. At the same time,

$$\begin{aligned}\mathbf{diam}_{d_1}(\mathcal{T}) &= \sup_{s,t\in\mathcal{T}} d_1(s,t)\\ &= \frac{1+3\left\|x_\star\right\|}{m}\sup_{s,t\in\mathcal{T}} d_{\mathsf{euc}}(s,t)\\ &\le \frac{1+3\left\|x_\star\right\|}{m}\cdot\sqrt{\sup_{u,\bar{u}\in\mathbb{S}^{d-1}}\left\|u-\bar{u}\right\|^2+\sup_{v,\bar{v}\in\mathbb{S}^{d-1}}\left\|v-\bar{v}\right\|^2}\\ &\le \frac{2\sqrt{2}(1+3\left\|x_\star\right\|)}{m},\end{aligned}$$

as expected. This completes the proof. ■

Proof of Claim 3. Proceeding as in the proof of Claim 2, we obtain

$$\begin{aligned}\gamma_2(\mathcal{T},d_2) &= \frac{1+\gamma}{\sqrt{m}}\gamma_2(\mathcal{T},\left\|\cdot\right\|)\\ &\lesssim \frac{1+\gamma}{\sqrt{m}}\int_0^\infty \sqrt{\log\mathcal{N}\left(\mathcal{T},u\right)}\mathrm{d}u\\ &= \frac{1+\gamma}{\sqrt{m}}\int_0^{\mathrm{diam}(\mathcal{T})} \sqrt{\log\mathcal{N}(\mathcal{T},u)}\mathrm{d}u\\ &\lesssim \frac{(1+\gamma)\sqrt{d}}{m}\int_0^1 \sqrt{\log\frac{1}{\varepsilon}}\mathrm{d}\varepsilon\\ &= (1+3\left\|x_\star\right\|)\sqrt{\frac{d}{m}}.\end{aligned}$$

Similarly, to control the diameter under d_2 we obtain

$$\begin{aligned}\mathbf{diam}_{d_2}(\mathcal{T}) &= \sup_{s,t\in\mathcal{T}} d_2(s,t)\\ &= \frac{1+3\left\|x_\star\right\|}{\sqrt{m}}\sup_{s,t\in\mathcal{T}} d_{\mathsf{euc}}(s,t)\\ &\le \frac{1+3\left\|x_\star\right\|}{\sqrt{m}}\cdot\sqrt{\sup_{u,\bar{u}\in\mathbb{S}^{d-1}}\left\|u-\bar{u}\right\|^2+\sup_{v,\bar{v}\in\mathbb{S}^{d-1}}\left\|v-\bar{v}\right\|^2}\\ &\le \frac{2\sqrt{2}(1+3\left\|x_\star\right\|)}{\sqrt{m}},\end{aligned}$$

as expected. This completes the proof. ■

Proof of Claim 4. Note that when $x = \mathbf{0}$, $y_i - h_i(0) = 1 - \exp\left(-\langle a_i, x_\star\rangle_+\right)$. As a result,

$$\begin{aligned} f(0) &= \frac{1}{m}\sum_{i=1}^{m} 1 - \exp\left(-\langle a_i, x_\star\rangle_+\right) \\ &= 1 - \frac{1}{m}\sum_{i=1}^{m} \exp\left(-\|x_\star\|\left\langle a_i, \frac{x_\star}{\|x_\star\|}\right\rangle_+\right) \\ &\overset{(d)}{=} 1 - \frac{1}{m}\sum_{i=1}^{m} \exp\left(-(\beta_i)_+\|x_\star\|\right), \qquad \beta_i \sim \mathcal{N}(0,1). \end{aligned} \tag{40}$$

We now argue the sum in (40) concentrates. To that end, we first calculate its expectation:

$$\begin{aligned} \mathbb{E}[f(0)] &= 1 - \mathbb{E}[e^{-\beta_+\|x_\star\|}] \\ &= 1 - \frac{1}{2}\left(1 + \exp\left(\frac{\|x_\star\|^2}{2}\right)\operatorname{erfc}\left(\frac{\|x_\star\|}{\sqrt{2}}\right)\right) \\ &= \frac{1}{2}\left(1 - \exp\left(\frac{\|x_\star\|^2}{2}\right)\operatorname{erfc}\left(\frac{\|x_\star\|}{\sqrt{2}}\right)\right). \end{aligned}$$

To show the sum concentrates, we use the Gaussian Lipschitz inequality. We have

$$\begin{aligned} &\frac{1}{m}\sum_{i=1}^{m}(1 - \exp(-\langle a_i, x_\star\rangle_+)) - \frac{1}{m}\sum_{i=1}^{m}(1 - \exp(-\langle \widetilde{a}_i, x_\star\rangle_+)) \\ &= \frac{1}{m}\sum_{i=1}^{m}\exp(-\langle \widetilde{a}_i, x_\star\rangle_+) - \exp(-\langle a_i, x_\star\rangle_+) \\ &\le \frac{1}{m}\sum_{i=1}^{m}|\langle a_i - \widetilde{a}_i, x_\star\rangle| \\ &\le \frac{\|x_\star\|}{\sqrt{m}}\left\|A - \widetilde{A}\right\|_{\mathsf{F}}, \qquad \text{where} \quad A = [a_1 \quad \dots \quad a_m]^{\mathsf{T}},\ \widetilde{A} := [\widetilde{a}_1 \quad \dots \quad \widetilde{a}_m]^{\mathsf{T}}, \end{aligned}$$

using the fact that the function $x \mapsto \exp(-x_+)$ is 1-Lipschitz. We deduce that $f(0)$ is Lipschitz with modulus $\frac{\|x_\star\|}{\sqrt{m}}$ with respect to the random vectors $\{a_i\}_{i=1,\dots,m}$. Invoking the Gaussian Lipschitz concentration inequality [29, Theorem 5.2.2] yields

$$\mathbb{P}\left(|f(0) - \mathbb{E}\left[f(0)\right]| \ge \|x_\star\|\sqrt{\frac{2d}{m}}\right) \le 2\exp\left(-d\right). \tag{41}$$

Finally, we bound $\mathbb{E}\left[f(0)\right]$ for $\|x_\star\| \ge 1$. Indeed, Lemma E.2 shows $\exp(\frac{u^2}{2})\operatorname{erfc}(\frac{u}{\sqrt{2}})$ is monotone decreasing on $u \in (0,\infty)$. As a result, we can lower bound the expectation by

$$\begin{aligned} \mathbb{E}\left[f(0)\right] &= \frac{1}{2}\left(1 - \exp\left(\frac{\|x_\star\|^2}{2}\right)\operatorname{erfc}\left(\frac{\|x_\star\|}{\sqrt{2}}\right)\right) \\ &\ge \frac{1}{2}\left(1 - \sqrt{\mathrm{e}}\cdot\operatorname{erfc}\left(\frac{1}{\sqrt{2}}\right)\right) \end{aligned}$$

$$\geq \frac{1}{2}(1 - 0.53)$$
$$= 0.235,$$

after numerically evaluating $\sqrt{\mathrm{e}} \cdot \mathrm{erfc}(1/\sqrt{2}) \leq 0.53$. Adjusting m so that $\sqrt{\frac{2d}{m}} \leq 0.035$ and plugging it into (41) yields the desired lower bound. ■

D.1. Proof of Lemma 2.1

Proof The main stepping stone to (8) is the following "decrease principle" (see [8, Theorem 3.2.8] for a version using the so-called *proximal* subdifferential):

Claim 5 (Decrease principle) *Let f be locally Lipschitz and fix $\rho, \mu > 0$. Suppose that*

$$z \in \mathcal{B}(x;\rho), \quad \zeta \in \partial f(z) \implies \|\zeta\| \geq \mu, \tag{42}$$

Then the following inequality holds:

$$\inf_{z \in \mathcal{B}(x;\rho)} f(z) \leq f(x) - \rho\mu. \tag{43}$$

Before proving Claim 5, we show how it implies the conclusion in (8). Indeed, suppose the conclusion were false; then, for some $x \in \mathcal{B}(x_\star; \|x_\star\|)$, there is $\rho > 0$ such that

$$\min\left(\|x - x_\star\|, \|x\|\right) > \rho > \frac{f(x) - f_\star}{\mu}.$$

From Lemma E.5 and $\|x\| \leq 2\|x_\star\|$ it follows that

$$\min\left(\|x - x_\star\|, \|x\|, \mathrm{dist}(x, \mathcal{B}^c(0; 3\|x_\star\|)\right) = \min(\|x - x_\star\|, \|x\|) > \rho > \frac{f(x) - f_\star}{\mu}. \tag{44}$$

At the same time, since $\min(\cdot)$ is associative, we have that

$$\min(\|x\|, \mathrm{dist}(x, \mathcal{B}^c(0; 3\|x_\star\|))) = \mathrm{dist}\left(x, (\mathcal{B}(0; 3\|x_\star\|) \setminus \{0\})^c\right).$$

Consequently, it follows that

$$\mathcal{B}(x;\rho) \subset \mathcal{B}(0; 3\|x_\star\|) \setminus \{0\}, \quad \|x - x_\star\| > \rho.$$

The second conclusion in the display above implies that $x \neq x_\star$. Therefore,

$$x \in \mathcal{B}(0; 3\|x_\star\|) \setminus \{0, x_\star\} \overset{\text{(Aiming)}}{\implies} \min_{v \in \partial f(x)} \|v\| \geq \mu.$$

As a result, invoking Claim 5, we obtain

$$0 \leq \inf_{\bar{x} \in \mathcal{B}(x;\rho)} f(\bar{x}) - f_\star \leq (f(x) - f_\star) - \rho\mu \overset{(44)}{<} 0, \tag{45}$$

which is a contradiction with our assumption that (8) fails; therefore, (8) must hold.

Proof [Proof of Claim 5] It remains to prove Claim 5. Since this is essentially the same as [8, Theorem 3.2.8], with the proximal subdifferential replaced by the Clarke subdifferential, it suffices to repeat its proof with a single modification: instead of applying the version of the mean-value inequality from [8, Theorem 3.2.6], we invoke [7, Theorem 4.1], which is valid for the Clarke subdifferential. ■

This completes the proof of the Lemma. ■

Appendix E. Technical results

Lemma E.1 *For $X \sim \mathcal{N}(0,1)$, we have that*

$$\mathbb{E}\left[e^{-cX}\mathbb{1}\{X \geq 0\}\right] = \frac{1}{2}\exp\left(\frac{c^2}{2}\right)\operatorname{erfc}\left(\frac{c}{\sqrt{2}}\right), \tag{46a}$$

$$\mathbb{E}\left[e^{-cX}X_+\right] = \frac{1}{2}\left(\sqrt{\frac{2}{\pi}} - c\exp\left(-\frac{c^2}{2}\right)\operatorname{erfc}\left(\frac{c}{\sqrt{2}}\right)\right). \tag{46b}$$

Proof The first expectation is the integral

$$\begin{aligned}
\mathbb{E}\left[e^{-cX}\mathbb{1}\{X \geq 0\}\right] &= \int_0^\infty \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2} - cx\right)\mathrm{d}x \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}}\exp\left(-\left(\frac{x^2}{2} + cx + \frac{c^2}{2}\right)\right)\exp\left(\frac{c^2}{2}\right)\mathrm{d}x \\
&= \exp\left(\frac{c^2}{2}\right)\int_0^\infty \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x+c)^2}{2}\right)\mathrm{d}x \\
&= \exp\left(\frac{c^2}{2}\right)\int_{\frac{c}{\sqrt{2}}}^\infty \frac{1}{\sqrt{\pi}}\exp\left(-z^2\right)\mathrm{d}z \qquad \left(z \leftarrow \frac{x+c}{\sqrt{2}}\right) \\
&= \exp\left(\frac{c^2}{2}\right)\frac{1}{2}\cdot\frac{2}{\sqrt{\pi}}\int_{\frac{c}{\sqrt{2}}}^\infty \exp\left(-z^2\right)\mathrm{d}z \\
&= \frac{1}{2}\exp\left(\frac{c^2}{2}\right)\operatorname{erfc}\left(\frac{c}{\sqrt{2}}\right),
\end{aligned}$$

using the definition of erfc in the last equality; this proves (46a). The second expectation is

$$\begin{aligned}
\mathbb{E}\left[e^{-cX}X_+\right] &= \int_0^\infty \frac{x}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2} - cx\right)\mathrm{d}x \\
&= \int_0^\infty \frac{(x+c)}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2} - cx\right)\mathrm{d}x - c\cdot\int_0^\infty \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2} - cx\right)\mathrm{d}x \\
&= \exp\left(\frac{c^2}{2}\right)\int_{\frac{c}{\sqrt{2}}}^\infty z\cdot\sqrt{\frac{2}{\pi}}\exp\left(-z^2\right)\mathrm{d}z - \frac{c}{2}\exp\left(\frac{c^2}{2}\right)\operatorname{erfc}\left(\frac{c}{\sqrt{2}}\right),
\end{aligned}$$

recognizing the integral from (46a) in the second term. Note that

$$\frac{\mathrm{d}}{\mathrm{d}z}e^{-z^2} = -2ze^{-z^2} \implies \int_0^\infty z\cdot\sqrt{\frac{2}{\pi}}e^{-z^2}\mathrm{d}z = -\sqrt{\frac{2}{\pi}}\frac{1}{2}\left\{e^{-z^2}\right\}_{\frac{c}{\sqrt{2}}}^\infty = \frac{1}{2}\exp\left(-\frac{c^2}{2}\right)\sqrt{\frac{2}{\pi}},$$

cancelling out the leading $\exp(c^2/2)$; this proves (46b). ∎

Lemma E.2 *The function $f(x) := \exp(x^2)\cdot\operatorname{erfc}(x)$ is monotone decreasing for $x \geq 0$.*

Proof The first derivative of f is equal to

$$f'(x) = 2x \cdot \exp(x^2)\mathrm{erfc}(x) - \frac{2}{\sqrt{\pi}}.$$

We now use the inequality $\mathrm{erfc}(x) \leq \frac{2e^{-x^2}}{\sqrt{\pi}(x+\sqrt{x^2+4/\pi})}$, which yields

$$\begin{aligned} f'(x) &\leq 2x \cdot \frac{2}{\sqrt{\pi}} \cdot \frac{1}{x + \sqrt{x^2 + 4/\pi}} - \frac{2}{\sqrt{\pi}} \\ &< \frac{4x}{\sqrt{\pi}} \cdot \frac{1}{2x} - \frac{2}{\sqrt{\pi}} \\ &\leq 0, \end{aligned}$$

which completes the proof of the claim. ∎

Lemma E.3 *The function $\varphi(x) := x \exp(x^2) \cdot \mathrm{erfc}(x)$ is monotone increasing for $x \geq 0$.*

Proof The first derivative of φ is

$$\varphi'(x) = \exp(x^2)(2x^2 + 1)\mathrm{erfc}(x) - \frac{2x}{\sqrt{\pi}}$$

Clearly, $\varphi'(0) > 0$. Now for any $x > 0$, we have

$$\begin{aligned} e^{x^2}\mathrm{erfc}(x) &\geq \frac{2}{\sqrt{\pi}(x + \sqrt{x^2 + 2})} \\ &= \frac{2}{\sqrt{\pi}} \frac{1}{x\left(1 + \sqrt{1 + \frac{2}{x^2}}\right)} \\ &\overset{(\sharp)}{>} \frac{2}{x\sqrt{\pi}} \frac{1}{2 + \frac{1}{x^2}} \\ &= \frac{2}{\sqrt{\pi}} \frac{1}{2x + \frac{1}{x}} \\ &= \frac{2x}{\sqrt{\pi}} \frac{1}{2x^2 + 1}, \end{aligned}$$

where $(\sharp)$ follows from $\sqrt{1+x} < 1 + \frac{x}{2}$, valid for any $x > 0$. Finally, multiplying with $(2x^2 + 1)$ and subtracting $2x/\sqrt{\pi}$ yields $\varphi'(x) > 0$ for any $x > 0$, completing the proof. ∎

Lemma E.4 *For any $x \geq 1$, the following bound holds:*

$$x \exp\left(\frac{x^2}{2}\right) \mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right) \leq \sqrt{\frac{2}{\pi}} \left(1 - \frac{1}{1 + \pi x^2}\right). \tag{47}$$

Proof Starting from the known inequality

$$\operatorname{erfc}(x) \leq \frac{2e^{-x^2}}{\sqrt{\pi}(x+\sqrt{x^2+\frac{4}{\pi}})},$$

we successively obtain

$$\begin{aligned} x\exp(x^2/2)\operatorname{erfc}(x/\sqrt{2}) &\leq \frac{2x}{\sqrt{\pi}}\frac{1}{\frac{x}{\sqrt{2}}+\sqrt{\frac{x^2}{2}+\frac{4}{\pi}}} \\ &= \frac{2\sqrt{2}}{\sqrt{\pi}}\frac{1}{1+\sqrt{1+\frac{8}{\pi x^2}}} \\ &\leq \frac{2\sqrt{2}}{\sqrt{\pi}}\frac{1}{2+\frac{2}{\pi x^2}} \\ &= \sqrt{\frac{2}{\pi}}\frac{1}{1+\frac{1}{\pi x^2}}, \end{aligned}$$

where the last inequality is due to the fact that

$$\sqrt{1+y^2} \geq 1+\frac{y^2}{4}, \quad \forall y \in [0,1].$$

Finally, we rewrite the last fraction as

$$\frac{1}{1+\frac{1}{\pi x^2}} = \frac{1+\frac{1}{\pi x^2}}{1+\frac{1}{\pi x^2}} - \frac{\frac{1}{\pi x^2}}{1+\frac{1}{\pi x^2}} = 1 - \frac{1}{\pi x^2+1}.$$

This completes the proof of (47). ∎

Lemma E.5 *Suppose that x satisfies $\|x-x_\star\|^2 < (1-\delta)\|x_\star\|^2$. Then*

$$\min\left(\|x-x_\star\|, \|x\|, \operatorname{dist}(x, \mathcal{B}^c(\mathbf{0}; 3\|x_\star\|))\right) \geq \min\{1, \frac{\delta}{2\sqrt{1-\delta}}\} \cdot \|x-x_\star\|.$$

Proof We have the following sequence of inequalities:

$$\begin{aligned} \operatorname{dist}(x, \mathcal{B}^c(\mathbf{0}; 3\|x_\star\|)) &= \left\|x - \operatorname{proj}_{\mathcal{B}^c(\mathbf{0};3\|x_\star\|)}(x)\right\| \\ &= \left\|x - x_\star + x_\star - \operatorname{proj}_{\mathcal{B}^c(\mathbf{0};3\|x_\star\|)}(x)\right\| \\ &\geq \left\|x_\star - \operatorname{proj}_{\mathcal{B}^c(\mathbf{0};3\|x_\star\|)}(x)\right\| - \|x-x_\star\| \\ &\geq \left\|x_\star - \operatorname{proj}_{\mathcal{B}^c(\mathbf{0};3\|x_\star\|)}(x_\star)\right\| - \|x-x_\star\| \\ &> \left\|x_\star - \operatorname{proj}_{\mathcal{B}^c(\mathbf{0};3\|x_\star\|)}(x_\star)\right\| - \|x_\star\| \\ &\geq 2\|x_\star\| - \|x_\star\| \\ &> \|x-x_\star\|, \end{aligned}$$

which follow from $\|x - \mathrm{proj}_C(y)\| \geq \|x - \mathrm{proj}_C(x)\|$, our assumptions on $\|x - x_\star\|$, and

$$\left\|x_\star - \mathrm{proj}_{\mathcal{B}^c(0;3\|x_\star\|)}(x_\star)\right\| \geq \left\|\mathrm{proj}_{\mathcal{B}^c(0;3\|x_\star\|)}(x_\star)\right\| - \|x_\star\| \geq 3\|x_\star\| - \|x_\star\| = 2\|x_\star\|.$$

This shows that $\|x - x_\star\| < \mathrm{dist}(x, \mathcal{B}^c(\mathbf{0}; 3\|x_\star\|))$. At the same time,

$$\begin{aligned}(1-\delta)\|x_\star\|^2 &\geq \|x - x_\star\|^2 \\ &= \|x\|^2 + \|x_\star\|^2 - 2\langle x, x_\star\rangle \\ &\geq \|x_\star\|^2 - 2\langle x, x_\star\rangle \\ \Leftrightarrow \left\langle x, \frac{x_\star}{\|x_\star\|}\right\rangle &\geq \frac{\delta}{2}\|x_\star\|.\end{aligned}$$

From the previous display and our assumption $\|x - x_\star\| \leq \sqrt{1-\delta}\,\|x_\star\|$, we deduce

$$\begin{aligned}\|x\| &= \sup_{u\in\mathbb{S}^{d-1}} \langle x, u\rangle \\ &\geq \left\langle x, \frac{x_\star}{\|x_\star\|}\right\rangle \\ &\geq \frac{\delta}{2}\|x_\star\| \\ &= \frac{\delta}{2}\frac{\|x_\star\|}{\|x - x_\star\|}\cdot\|x - x_\star\| \\ &\geq \frac{\delta}{2\sqrt{1-\delta}}\|x - x_\star\|,\end{aligned}$$

which completes the proof of the claim. ■

Claim 6 *For any x satisfying $\|x - x_\star\| \leq (1-\delta)\|x_\star\|$, we have $\langle x, x_\star\rangle \geq \delta\|x_\star\|^2$.*

Proof The proof follows from the following inequality:

$$\langle x, x_\star\rangle = \langle x - x_\star, x_\star\rangle + \|x_\star\|^2 \geq \|x_\star\|\left(\|x_\star\| - \|x - x_\star\|\right) \geq \delta\|x_\star\|^2.$$

using the bound on $\|x - x_\star\|$ in the last step. ■