A Fast Lewis-weight Interior Point Method for Linear Optimization over Product-Domain

Peiyuan Zhang
UW-Madison
pzhang287@wisc.edu

Stephen Wright
UW-Madison
swright@cs.wisc.edu

November 2026

Abstract

We study the following linear optimization problem over product domains:

$$\min_{\substack{x\in K_1\times\cdots\times K_m \\ A^\top x=b}} c^\top x,$$

where $\mathbf{A}\in\mathbb{R}^{n\times d}$, $b\in\mathbb{R}^d$, $c\in\mathbb{R}^n$, each $K_i\subseteq\mathbb{R}^{n_i}$ is a compact convex set of constant dimension, and $\sum_i n_i = n$. We assume that each K_i admits an efficiently computable self-concordant barrier. This framework generalizes linear programming and includes a broad class of empirical risk minimization problems as special cases.

We give a randomized interior-point method that computes a high-accuracy solution in total time $\widetilde{O}(nd+d^{5/2})$, using only $\widetilde{O}(\sqrt{d})$ iterations. This matches the best known running time for scalar linear programming in the tall dense regime. Compared with the previous $\widetilde{O}(nd+d^6\sqrt{n})$ bound for general product-domain optimization, our result removes the $\sqrt{n}$ dependence from the computational overhead, reducing the threshold for nearly input-linear time from $n\geq d^{10}$ to $n\geq d^{3/2}$.

Our algorithm extends the Lee–Sidford Lewis-weight framework beyond linear programming to general constant-dimensional convex domains. The main technical ingredients include block Lewis weights, randomized smoothing of inverse Hessians, and a new Bregman-divergence-based centrality measure. Together with efficient dynamic data structures, these techniques allow us to match the fastest known running time for scalar LP under ordinary self-concordance, without the higher-order self-concordance assumption used in previous fast implementations.

Contents

1 Introduction **1**
- 1.1 Our Results . . . 2
- 1.2 Related Work . . . 3

2 Overview **3**
- 2.1 Block Lewis-weight Central Path Following . . . 3
- 2.2 Removing HOSC from Fast IPM Implementation . . . 6
- 2.3 Smoothed Scaling . . . 6
- 2.4 Bregman Divergence and Potential . . . 7
- 2.5 Implementation and Data Structures . . . 9

3 Preliminaries **10**
- 3.1 Notations . . . 10
- 3.2 Self-concordant Barriers . . . 11
- 3.3 Fenchel Conjugate and Bregman Divergence . . . 11
- 3.4 Randomized smoothing . . . 12

4 Block Lewis-weight IPM **12**
- 4.1 Supporting and Technical Lemma . . . 16
- 4.2 Stability of Variables . . . 22
- 4.3 Bounds of Bregman Metric . . . 25
- 4.4 Potential Descent Analysis . . . 30
- 4.5 Feasibility and Sampling Scheme . . . 37

5 Overview of Data Structures **38**
- 5.1 Dual and Gradient Maintenance . . . 38
- 5.2 Smoothing Scaling Maintenance . . . 39
- 5.3 Lewis Weight Maintenance . . . 41
- 5.4 Sampling and Inverse Maintenance . . . 43

6 Implementation of Path Following Algorithm **44**
- 6.1 Initial and Final Processing . . . 44
- 6.2 Nearby Sequence Properties . . . 45
- 6.3 Short-Step Implementation . . . 45
- 6.4 Path Following Analysis . . . 49

A Basic Mathematical Facts and Proofs **55**
- A.1 Proofs involving Self-concordance, Fenchel Conjugate and Bregman Divergence . . . 55
- A.2 Projection Matrix facts . . . 56
- A.3 Block Lewis-weight Facts . . . 58

B Proofs of IPM Analysis **60**
- B.1 Stability of x, s and $\mathbf{C}$ . . . 60
- B.2 Stability of τ . . . 63
- B.3 Stability of μ . . . 71

B.4 Proof of Approximate Feasibility 75
B.5 Sampling Scheme Construction 76

C Proofs of Data Structures **77**
C.1 Dual Maintenance 77
C.2 Gradient Maintenance 80
C.3 Leverage Score Maintenance 83
C.4 Samping and Inverse Maintenance 88

D Remaining Proofs for Section 6 **89**
D.1 Initial and Final Processing 89
D.2 Proof of Nearby Sequence 91

1 Introduction

Empirical risk minimization (ERM) is a fundamental problem in statistical learning. Given convex loss functions $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ and matrices $\mathbf{A}_i \in \mathbb{R}^{n_i \times d}$, one seeks to solve

$$\min_{y \in \mathbb{R}^d} \sum_{i=1}^{m} f_i(\mathbf{A}_i y - c_i).$$

Many important problems, including linear and ℓ_p regression, LASSO, logistic regression, support vector machines, and quantile regression, can be formulated in this framework [LSZ19, LPT+25]. Here, we focus on the *tall dense regime*, where the total number of observations $n = \sum_{i=1}^{m} n_i$ is much larger than the number of features d, and the input size is $\Theta(nd)$. A central algorithmic goal is to obtain high-accuracy solutions with only logarithmic dependence on $1/\varepsilon$, while keeping the running time close to the dense input size nd. In particular, can we achieve a running time comparable to that of the fastest algorithms for linear programming (LP), with only a small overhead beyond reading the input?

By taking convex conjugates and applying minimax duality, followed by an epigraph reformulation, ERM can be expressed as a special case of the following optimization problem, referred to as *linear optimization over product domains*:

$$\min_{x \in \prod_{i=1}^{m} K_i,\ \mathbf{A}^\top x = b} c^\top x, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^d$, $c \in \mathbb{R}^n$, and each $K_i \subseteq \mathbb{R}^{n_i}$ is a convex set. We study this class of problems in this work. More precisely, we assume that each K_i has constant dimension $n_i \le n_{\max} = O(1)$ and is equipped with a 1-self-concordant barrier ϕ_i.

When all K_i are one-dimensional intervals, (1) reduces to LPs. There is a long line of work on solving LPs with interior-point methods (IPMs), which provide a natural benchmark for the running times we seek. Since the development of polynomial-time IPMs [Kar84, Vai89], classical path-following algorithms have achieved $\widetilde{O}(\sqrt{n})$ iterations [Ren88]. For tall instances with $n \gg d$, Lee and Sidford [LS14, LS19] introduced Lewis-weight central paths that reduce the iteration complexity to $\widetilde{O}(\sqrt{d})$. Subsequent advances [vdBLSS20, vdBLN+20] developed dynamic data structures that enable efficient implementations of the Lee–Sidford approach. The fastest running time along this line for LP is $\widetilde{O}(nd + d^{2.5})$, obtained in [VDBLL+21].

In the general setting (1), the blocks may be nonpolyhedral, and the Lewis-weight IPM techniques developed for LPs do not directly extend to arbitrary self-concordant barriers. In particular, unlike in LP, where explicit logarithmic barriers are available, we assume access only to black-box barriers for the sets K_i, with no additional structure beyond self-concordance. Achieving a comparable result in this general setting has therefore proven considerably more challenging. Lee, Song, and Zhang [LSZ19] focused on the square setting ($n = \Theta(d)$) and gave an IPM running in matrix multiplication time $n^{\omega+o(1)}$. More recently, [LPT+25] introduced a novel adaptive matrix sparsification method and obtained the first fast algorithm for sufficiently tall dense ERM, with running time $\widetilde{O}(nd + d^6\sqrt{n})$. However, this running time becomes nearly linear in the input size only for extremely tall instances with $n \ge d^{10}$, leaving a substantial gap to the $\widetilde{O}(nd + d^{2.5})$ bound for LPs. This gap is structural: unlike the Lee–Sidford approach, their algorithm requires $\widetilde{O}(\sqrt{n})$ rather than $\widetilde{O}(\sqrt{d})$ iterations. Even if each iteration could be implemented in $O(d^2)$ time, the resulting $d^2\sqrt{n}$ term would already exceed the LP benchmark in the regime $d \ll n \ll d^2$. Improving the per-iteration

implementation alone cannot match the fastest LP algorithms throughout the tall regime; reducing the iteration complexity is therefore essential. However, extending the Lee–Sidford framework beyond LP to general self-concordant barriers for multidimensional convex sets K_i remains an open problem, as explicitly noted in [LPT+25]. Thus, we are led to the central question of our work:

Can the Lee–Sidford IPM be extended to product-domain programs with general self-concordant barriers, achieving the $\widetilde{O}(\sqrt{d})$ iteration complexity and $\widetilde{O}(nd + d^{2.5})$ running time of the fastest LP algorithms?

1.1 Our Results

We answer this question affirmatively by devising a Lewis-weight-based IPM for linear optimization over product domains in (1), matching the running time of the fastest scalar LP algorithm in [VDBLL+21] when each K_i has constant dimension.

Theorem 1 (Main result, informal)**.** *Consider (1). Suppose that $n_i = O(1)$ for any $i \in [m]$ and each K_i is equipped with a 1-self-concordant barrier, with constant-time oracle access to the value, gradient, and Hessian of ϕ_i, and to the value and gradient of its Fenchel conjugate[1]. For any $\delta > 0$, there exists an algorithm that outputs x such that $x \in K := \prod_{i=1}^m K_i$, $\mathbf{A}^\top x = b$, and*

$$c^\top x \le \min_{\substack{x \in K \\ \mathbf{A}^\top x = b}} c^\top x + \delta$$

in time $\widetilde{O}(nd + d^{2.5})$, where $\widetilde{O}(\cdot)$ hides logarithmic factors and polynomial dependence on the block size.

Our result establishes for the first time that fast Lee–Sidford path following extends to general constant-dimensional product domains equipped with black-box self-concordant barriers. Compared with [LPT+25], our result reduces the number of iterations from $O(\sqrt{n})$ to $\widetilde{O}(\sqrt{d})$ and the overhead from $\widetilde{O}(d^6\sqrt{n})$ to $\widetilde{O}(d^{5/2})$, independent of n. The $\sqrt{n}$ dependence is removed at its source, the iteration count, lowering the threshold for nearly linear time from $n \ge d^{10}$ to $n \ge d^{3/2}$.

Moreover, our work is not merely a blockwise extension of the Lee–Sidford approach: it also provides a new analysis for fast IPMs that eliminates the high-order self-concordance (HOSC) condition and requires only ordinary self-concordance. This is particularly important because breaking the d^3 barrier in the fastest LP implementations requires sampling the primal updates. The analysis in [VDBLL+21] relies on HOSC to obtain a sharp bound for the first-moment error arising from the sampled primal updates. For scalar LPs, the standard logarithmic barrier already satisfies HOSC, but this property is not guaranteed for general barrier functions on multidimensional blocks K_i. Our analysis avoids the additional assumption but still achieves the same running time as [VDBLL+21]. Our approach combines randomized smoothing of Hessian-dependent quantities with a Bregman-divergence-based centrality measure, allowing us to control the stochastic errors under ordinary self-concordance. We therefore broaden the applicability of the fastest known Lewis-weight IPM techniques, and we hope the ideas will be useful in other theory and optimization settings.

[1]The Fenchel conjugates are used only to evaluate our Bregman potential and its gradient. Under our centrality conditions, they can also be approximated efficiently using local Newton iterations with primal barrier oracles. We assume exact access for simplicity and omit the additional approximation analysis.

1.2 Related Work

Fast interior-point methods. Recent progress in high-accuracy interior-point methods has been driven by Lewis-weight reweighting, robust and stochastic path following, and efficient implementations [LS14,LS19,CLS19,LSZ19,vdB20,vdBLSS20,SY20,JSWZ21,vdBLN+20,VDBLL+21]. Related techniques have led to faster algorithms for semidefinite programming and sum-of-squares optimization [JKL+20,HJS+22,JNW22]. Interior-point methods and closely related optimization frameworks have also played a central role in advances for bipartite matching, maximum flow, minimum-cost flow, and generalized flow [Mad13, CMSV17, LS20, AMV21, CKL+22, vdBCP+23, JKLS25].

Dynamic data structures. Efficient maintenance of matrix inverses and linear-system solvers under low-rank or slowly varying updates has been studied extensively [San04, LS15, vdBNS19, vdB20, JSWZ21, vdB21]. Related developments include dynamic linear-system solutions and least-squares regression [CLS19, LSZ19, vdBLSS20, JPW23, AJK25], as well as maintaining approximate primal-dual variables, leverage scores, and adaptive spectral sparsifiers [vdBLSS20, vdBLN+20, VDBLL+21, LPT+25]. These algorithms build on a broad literature on sketching, subspace embeddings, and heavy-hitter data structures [JL84, KNPW11, Pag13, NN13, KN14, LNNT16, PSW17, CJN18, NS19, NSW19]. Numerical stability and bit complexity of inverse maintenance and dynamic matrix expressions have also been investigated [GPV23, AvdBGZ24].

Lewis weights and block-norm sparsification. Lewis weights have been extensively studied for ℓ_p sampling, matrix approximation, and efficient high-precision computation [CP15, FLPS22, AGS24, GSZ26]. Related work on sparsifying sums of norms and generalized linear models develops sampling and geometric approximation techniques for structured convex objectives [JLLS23, JLLS24]. Block Lewis weights and their associated change-of-measure geometry have been further developed for approximating matrix block norms [MO25], with subsequent applications to group distributionally robust regression and algorithms for computing block Lewis weights [MP26, Pad26]. Our setting instead concerns regularized, Hessian-preconditioned block Lewis weights that evolve along an interior-point trajectory, requiring stability under randomized updates and compatibility with the Newton geometry.

2 Overview

We provide an overview of the main techniques of our algorithm and analysis.

2.1 Block Lewis-weight Central Path Following

We restate our target problem as follows:

$$\min_{x \in K,\, \mathbf{A}^\top x = b} c^\top x \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^d$, $c \in \mathbb{R}^n$, and $K := \prod_{i=1}^m K_i$ is a product of m convex sets $K_i \subset \mathbb{R}^{n_i}$, with $\sum_{i \in [m]} n_i = n$ and $n_i = O(1)$. Let $x_i \in \mathbb{R}^{n_i}$ be the i-th block of $x = (x_1, \ldots, x_m) \in \mathbb{R}^n$ corresponding to K_i. Each set K_i is equipped with a 1-self-concordant (SC) barrier function $\phi_i : \mathrm{int}(K_i) \to \mathbb{R}$. The value, gradient, and Hessian of each ϕ_i can be evaluated in constant time.

A standard interior-point method (IPM) tracks the solutions of the following barrier-regularized problem as the path parameter $\rho > 0$ varies:

$$x(\rho) = \operatorname{argmin}_{\mathbf{A}^\top x = b} \; c^\top x + \rho \sum_{i=1}^{m} w_i \phi_i(x_i)$$

for a chosen weight vector $w \in \mathbb{R}^m_{>0}$. Let $\Phi(x) = \sum_{i=1}^m \phi_i(x_i)$ be the unweighted sum of barrier functions and $\mathbf{W} = \mathsf{diag}_{\mathrm{G}}(w)$, where $\mathsf{diag}_{\mathrm{G}}$ lifts a vector $w \in \mathbb{R}^m$ to a block-diagonal matrix whose i-th diagonal block is $w_i I_{n_i}$. For a fixed weight vector w, the first-order optimality conditions for the barrier-regularized problem can be written as follows:

$$\mathbf{A}^\top x = b, \qquad \mathbf{A}y + s = c, \qquad s + \rho \mathbf{W} \nabla \Phi(x) = 0, \tag{3}$$

where $y \in \mathbb{R}^d$ is the Lagrange multiplier and $s \in \mathbb{R}^n$ is the dual slack. However, computing the exact central point for every ρ is not computationally feasible. Instead, we track an approximate path and allow $s + \rho \mathbf{W} \nabla \Phi(x)$ to be nonzero. At each iteration, we decrease the path parameter ρ and update the primal-dual variables such that the centrality violation remains small. Moreover, it suffices to maintain the pair (x, s) explicitly, since the multiplier y can be recovered from dual feasibility.

Simply setting $w_i = 1$ recovers the classical self-concordant path-following method of Renegar [Ren88], which requires $\widetilde{O}(\sqrt{n})$ iterations. This approach is undesirable for the tall regime where $n \gg d$. For standard *scalar* LPs, where each x_i is one-dimensional and ϕ_i is the standard logarithmic barrier, Lee and Sidford [LS14,LS19] overcame this limitation by allowing w to depend on x through the ℓ_p-Lewis weights, effectively reducing the iteration complexity to $\widetilde{O}(\sqrt{d})$. Intuitively, Lewis weights reshape the local geometry by measuring the importance of each barrier in the d-dimensional space. In particular, their total mass is $O(d)$ rather than $O(n)$, allowing the weighted path-following analysis to depend on the smaller dimension factor. Moreover, these IPMs are also *robust* in the sense that they can be implemented using approximations to the primal-dual variables x, s and weights w. The latest result in the Lee–Sidford line, [VDBLL+21], uses regularized Lewis weights of the preconditioned matrix $\nabla^2 \Phi(x)^{-\frac{1}{2}} \mathbf{A}$, and shows that the IPM can be implemented in total time $\widetilde{O}(nd + d^{2.5})$.

Our convergence analysis follows the standard robust path-following framework. For each block $i \in [m]$, we define the following centrality residual

$$\mu_i := \frac{s_i}{\rho w_i} + \nabla \phi_i(x_i) \in \mathbb{R}^{n_i}$$

and measure its deviation through a blockwise centrality metric Γ_i. We replace the usual local quadratic metric with a Bregman-based metric, as discussed in a later subsection. In particular, we measure the centrality through the exponential potential

$$\Psi(\mu, x) \; := \; \sum_{i=1}^{m} \exp(\lambda \Gamma_i), \qquad \lambda = \Theta(\log m), \tag{4}$$

where λ is a sufficiently large penalty parameter. The algorithm uses the gradient of this potential to construct approximate projected Newton-type primal-dual updates. By maintaining the bound $\Psi(\mu, x) \le \mathrm{poly}(m)$ throughout the iterations, the analysis ensures that the iterates remain close

to the weighted central path as ρ decreases. This approach gives $O(\sqrt{d})$ short steps when Lewis weights are used.

Our proposed path-following framework for the convex program in (2) largely follows the robust Lee–Sidford approach. In contrast, the recent algorithm of [LPT+25] avoids the need to maintain Lewis weights, at the cost of $\widetilde{O}(\sqrt{n})$ IPM iterations and additional overhead. Reducing the iteration complexity to $\widetilde{O}(\sqrt{d})$ comes at a cost, as extending the Lee–Sidford framework to general product domains brings additional difficulties. Unlike the scalar setting, each barrier ϕ_i now corresponds to an entire block of n_i coordinates while carrying only one weight w_i. Consequently, scalar Lewis weights no longer directly capture the contribution of each barrier. To overcome this, we introduce regularized *block* ℓ_p Lewis weights with $p = 1 - \frac{1}{\log(m)}$, defined as the solution to

$$w = \mathrm{tr}_{\mathrm{G}}\left(\mathbf{P}(\mathbf{C}\mathbf{W}^{1/2-1/p}\mathbf{A})\right) + \tfrac{d}{m}\mathbf{1}_m \tag{5}$$

where $\mathbf{P}(\cdot)$ denotes the orthogonal projector onto the column space of its argument, tr_{G} returns the vector of traces of the diagonal blocks under the block partition induced by x, and $\mathbf{C}$ is a block-diagonal preconditioner matrix. When all $n_i = 1$, it recovers the regularized scalar ℓ_p Lewis weights.

A key difficulty in extending the Lewis-weight approach to multidimensional blocks is to ensure the stability of the block Lewis weights: when the preconditioner $\mathbf{C}$ changes, we need to guarantee that the corresponding weights w do not change too much. For scalar LPs, this is relatively straightforward, since multiplicative changes in scalar preconditioners directly translate into comparable changes in the Lewis weights. Even for multidimensional blocks, the natural choice $\mathbf{C} = \nabla^2\Phi(x)^{-\frac{1}{2}}$ admits the required stability: Ordinary self-concordance provides direct spectral comparisons of the underlying Hessian geometry, without introducing an additional loss from noncommutativity. However, the fast randomized implementation requires stronger stability properties and necessitates a different preconditioner, as explained in the next subsection. This makes the stability analysis more demanding. Because of noncommutativity, small multiplicative changes in general SPD preconditioners need not translate into comparably small changes in the Lewis weights. A condition-number dependence might arise from anisotropy within each block. Another natural alternative is to use spherical scaling within each block, which avoids the noncommutativity issue by ignoring the internal anisotropy. However, each IPM iteration requires solving a weighted linear system to compute the Newton step, and the resulting Lewis weights are no longer comparable to the leverage scores of this system. This prevents the desired inverse-maintenance framework from applying.

To overcome these difficulties, we construct SPD matrices $\mathbf{S}_i$ approximating $\nabla^2\phi_i(x_i)^{-1}$ and take their positive definite square roots $\mathbf{C}_i = \mathbf{S}_i^{1/2}$ as the preconditioners. This is because the block leverage scores depend on $\mathbf{C}_i$ only through its square $\mathbf{C}_i^2 = \mathbf{S}_i$. We explain the construction of $\mathbf{S}_i$ in a later subsection. This allows us to control the stability of the block Lewis weights directly through the squared scaling $\mathbf{S}_i$, without introducing a condition-number dependence, while preserving their compatibility with the Newton geometry. The resulting framework extends the basic Lewis-weight path-following approach to general block-structured barriers. The remaining challenge is to match the running time of the fastest scalar LP algorithms. As we explain next, the fastest subcubic scalar LP implementations rely on higher-order self-concordance, which cannot be assumed for the general barriers considered here.

2.2 Removing HOSC from Fast IPM Implementation

Our goal is not only to extend Lewis-weight path following to general block-structured problems, but also to achieve a *subcubic* dependence on d beyond the input-size cost, matching the fastest scalar result. A major obstacle to fast IPM implementation arises from approximate linear-system solves, which make exact primal feasibility expensive to preserve. In [vdBLSS20], a direct correction step to restore approximate primal feasibility leads to an $O(d^3)$ total computational overhead. This cubic bottleneck was overcome by a sampling-based stochastic primal update, first introduced in [vdBLN+20] and subsequently employed in [VDBLL+21] to achieve an $\widetilde{O}(nd+d^{2.5})$ total running time for scalar LP.

However, this faster implementation brings additional requirements on the IPM analysis itself. In particular, the IPM analysis and the implementation are interwoven, as the diagonal sampling is incorporated into the primal update. The randomness in the update of x propagates to the matrix preconditioner, Lewis weights, and other Hessian-dependent quantities. A tension arises: Sparse randomized updates are essential for computational efficiency, while the IPM analysis requires sufficiently sharp bounds on the expected relative changes of these quantities. To obtain sharp bounds and break the cubic running-time barrier, [VDBLL+21] relies heavily on the high-order self-concordance (HOSC) assumption.

Formally, HOSC additionally requires a scalar 1-self-concordant barrier ϕ to satisfy $|\phi''''(x)| \leq 6|\phi''(x)|^2$. This fourth-derivative condition controls the second-order remainder in the variation of the inverse-Hessian square-root scaling, allowing its expected change to retain the sharp first-order bound despite randomized updates. In the scalar LP analysis, the condition automatically holds for the standard logarithmic barrier, as its higher-order derivatives are reciprocals of powers in x. The situation is less clear when K_i is multidimensional. A direct extension could impose analogous bounds on every directional higher-order derivative. However, such additional regularity is not guaranteed by ordinary self-concordance. More importantly, in the general convex program (2), the barrier functions are given through black-box oracles. Rather than extending HOSC to the block setting, we therefore seek to remove this requirement entirely while retaining the running time of the fastest scalar LP algorithms.

Removing HOSC presents two main difficulties. First, ordinary self-concordance does not guarantee the required second-order smoothness of the Hessian-based preconditioner. We address this by smoothing the inverse Hessian. Second, the conventional quadratic centrality metric involves a moving inverse Hessian, whose expected variation is similarly difficult to control. Since the underlying local geometry cannot be freely modified, we instead adopt a Fenchel–Bregman divergence that preserves this geometry while avoiding the problematic Hessian variation. We develop these two ideas in the following subsections.

2.3 Smoothed Scaling

We first consider a natural choice of preconditioners $\mathbf{C}_i$ for the block Lewis weights. A direct extension of the scalar case in [VDBLL+21] would set $\mathbf{C}_i = \nabla^2\phi_i(x_i)^{-1/2}$. To control the evolution of the Lewis weights, we need a sharp bound on the expected relative change of $\mathbf{C}_i$, of order $(1+o(1))\beta$, where β is the short-step size. However, ordinary self-concordance does not guarantee such a bound. Suppose the primal variable x_i is updated by $\delta_{x,i}$. We would need a second-order expansion of the form

$$\mathbb{E}[\mathbf{C}_i(x_i+\delta_{x,i}) - \mathbf{C}_i(x_i)] = \mathrm{D}\mathbf{C}_i(x_i)[\mathbb{E}[\delta_{x,i}]] + O\left(\mathbb{E}[\|\delta_{x,i}\|^2]\right). \tag{6}$$

Such an estimate requires Lipschitz continuity of the derivative of $\mathbf{C}_i$, a property not guaranteed by ordinary self-concordance. This point accounts for the reliance of previous analyses on HOSC.

A standard way to obtain the missing second-order smoothness is *randomized smoothing*. For a Lipschitz function f, consider its convolution with the uniform distribution on a unit Euclidean ball:

$$f_\sigma(x) = \mathbb{E}_{u\sim\mathrm{Unif}(B)}[f(x+\sigma u)].$$

Standard smoothing theory provides a quadratic bound on the first-order Taylor remainder of $f_\sigma(x)$, as required by our analysis.

However, directly smoothing the inverse Hessian square root does not readily yield the squared-preconditioner stability bounds needed for our block Lewis-weight analysis, as discussed in the previous subsection. Thus, we smooth the *inverse Hessian* and then take its positive definite square root, that is,

$$\widetilde{\mathbf{S}}_i(x_i, M_i) := \mathbb{E}_{u\sim\mathrm{Unif}(B_{n_i})}\left[\nabla^2\phi_i\left(x_i + \sigma M_i^{-1/2}u\right)^{-1}\right], \qquad \widetilde{\mathbf{C}}_i(x_i, M_i) := \widetilde{\mathbf{S}}_i(x_i, M_i)^{1/2}.$$

Our preconditioner is obtained by taking $M_i = \nabla^2\phi_i(x_i)$. This construction has two important features. First, smoothing the inverse Hessian before taking its square root allows us to control the squared preconditioner directly, without introducing a dependence on the condition number. Second, smoothing in the local Hessian metric ensures that all queried points remain inside K_i, by the Dikin ellipsoid property. Moreover, ordinary self-concordance guarantees that the smoothed inverse Hessian remains a $(1\pm O(\sigma))$-spectral approximation of the original inverse Hessian, thereby preserving the desired local geometry.

Most importantly, the smoothed preconditioner inherits the second-order smoothness needed for randomized updates. In particular, the Lipschitz continuity of its derivative allows us to obtain, for a fixed local metric M_i,

$$\mathbb{E}\left[\widetilde{\mathbf{C}}_i(x_i+\delta_{x,i}, M_i) - \widetilde{\mathbf{C}}_i(x_i, M_i)\right] = \mathrm{D}_x\widetilde{\mathbf{C}}_i(x_i, M_i)[\mathbb{E}[\delta_{x,i}]] + O\left(\frac{\mathbb{E}[\|\delta_{x,i}\|_{M_i}^2]}{\sigma}\right),$$

where the remainder is measured in the relative local operator norm. Crucially, the first-order term depends only on the expected displacement, while the randomness contributes through a quadratic error. This allows us to obtain the desired $(1+o(1))\beta$ first-order stability under randomized short steps, previously achieved using HOSC, while also ensuring the matrix stability required by block Lewis weights.

Finally, $\widetilde{\mathbf{C}}$ describes an ideal preconditioner defined through a continuous expectation, which cannot be evaluated exactly in the barrier-oracle model. For the actual algorithm, we replace this expectation with a finite-sample approximation, allowing the smoothed preconditioner to be computed and maintained efficiently. We return to this implementation in the final subsection.

2.4 Bregman Divergence and Potential

To track the weighted central path, the IPM maintains a small centrality residual rather than computing the exact central point at every iteration. For each block $i \in [m]$, recall that the centrality residual is defined as

$$\mu_i = \frac{s_i}{\rho w_i} + \nabla\phi_i(x_i), \tag{7}$$

which vanishes on the exact central path. In the formal analysis, we use τ to denote the regularized block Lewis weights, reserving w for a general weight vector. A natural measure of centrality is the squared local norm

$$\gamma_i := \|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}},$$

which is widely used in potential-reduction analyses of both scalar and block IPMs. A crucial difficulty is that the local geometry itself changes under randomized primal updates. In particular, the expected variation of γ_i takes the form

$$\mathbb{E}[\delta_{\gamma,i}] = 2\big\langle \nabla^2\phi_i(x_i)^{-1}\mu_i, \mathbb{E}[\delta_{\mu,i}]\big\rangle + \big\langle \mu_i, \mathbb{E}[\delta(\nabla^2\phi_i(x_i)^{-1})]\mu_i\big\rangle + \text{higher-order terms}.$$

While the first term captures the desired descent direction, the second involves the variation of the inverse Hessian and requires sharp higher-order control that ordinary self-concordance does not provide. This resembles the preconditioning difficulty discussed in the previous subsection, but with an important distinction: the preconditioner is an algorithmic choice that can be modified, whereas the local Hessian is intrinsic to the barrier geometry and cannot simply be replaced by a smoother proxy.

Our approach is instead to change the centrality measure while preserving the underlying local geometry. Specifically, we use the Bregman divergence induced by the Fenchel conjugate $\phi_i^\star$:

$$\Gamma_i(\mu_i, x_i) = 2D_{\phi_i^\star}\left(\nabla\phi_i(x_i) - \mu_i, \nabla\phi_i(x_i)\right). \tag{8}$$

Within an ϵ-neighborhood of the central path, this metric is locally equivalent to the squared residual norm, satisfying $\Gamma_i = (1 \pm O(\epsilon))\gamma_i$. Thus, it retains the same local geometry while admitting a more favorable variation formula. To see this, let $v_i = \nabla\phi_i^\star(\nabla\phi_i(x_i) - \mu_i) - x_i$. The expected variation of Γ_i can be written as

$$\mathbb{E}[\delta_{\Gamma,i}] = -2\big\langle v_i, \mathbb{E}[\delta_{\mu,i}]\big\rangle + 2\big\langle \nabla^2\phi_i(x_i)v_i + \mu_i, \mathbb{E}[\delta_{x,i}]\big\rangle + \text{higher-order terms}.$$

Crucially, v_i agrees with $-\nabla^2\phi_i(x_i)^{-1}\mu_i$ to first order, while

$$\big\|\nabla^2\phi_i(x_i)v_i + \mu_i\big\|_{\nabla^2\phi_i(x_i)^{-1}} = O\left(\|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}}\right).$$

Hence, the first term preserves the desired descent direction, whereas the contribution of the changing primal geometry is only second-order in the centrality residual. Unlike the quadratic metric, the Bregman divergence avoids explicitly differentiating the inverse Hessian, thereby removing the problematic first-order geometric variation.

Using this metric, we define the centrality potential

$$\Psi(\mu, x) = \sum_{i=1}^m \psi(\Gamma_i(\mu_i, x_i)) = \sum_{i=1}^m \exp\left(\lambda\Gamma_i(\mu_i, x_i)\right). \tag{9}$$

The resulting variation structure allows us to establish the expected potential decrease under randomized short steps using only ordinary self-concordance, while retaining the centrality guarantees needed for robust path following.

Although the Bregman divergence is nonlinear, its variation can still be controlled using essentially the same local-norm arguments as in the standard potential analysis. In particular, the Bregman divergence admits the integral representation

$$\Gamma_i(\mu_i, x_i) = 2\int_0^1 (1-t)\,\|\mu_i\|^2_{\nabla^2\phi_i^\star(\nabla\phi_i(x_i) - t\mu_i)}\,dt.$$

Within an ϵ-neighborhood of the central path, self-concordance ensures that the dual Hessian along this segment remains close to $\nabla^2\phi_i(x_i)^{-1}$. This representation allows us to control the nonlinear remainder using local quadratic estimates. Together with the preceding variation formula, we establish the first-order and second-moment bounds needed for the exponential-potential analysis, without relying on the HOSC assumption used in [VDBLL+21].

2.5 Implementation and Data Structures

Our robust path-following analysis allows us to work with approximations of the primal-dual variables and Lewis weights, without explicitly updating their exact values at every iteration. To achieve the desired running time, we follow the dynamic data structure framework of [vdBLSS20, VDBLL+21, LPT+25], which combines lazy updates, randomized sketching, and amortized analysis to avoid repeatedly processing all coordinates. In particular, these structures detect significant changes in the maintained quantities and refresh their approximations only when necessary. As in previous fast IPMs, the randomized structures must remain reliable under adaptive update sequences, since future iterates may depend on earlier randomized computations.

There are two main challenges in adapting this framework to our setting. First, the existing structures are primarily designed for diagonal scalings, whereas our block preconditioner consists of full matrices. Such updates can change row directions rather than merely rescale them, requiring new blockwise detection and verification procedures. Second, our smoothed preconditioner is defined through a continuous expectation and must be replaced by an efficiently maintainable finite-sample approximation. Most of these data structures are direct block extensions of their counterparts in [VDBLL+21]. Since each block has constant dimension $n_i = O(1)$, the corresponding correctness arguments largely carry over, and block operations incur only constant-factor overhead.

Finite-sum smoothing. We implement the smoothed preconditioner by replacing the expectation of inverse Hessians with an empirical average, followed by taking its positive definite square root. Matrix concentration bounds provide the required relative spectral approximation, while the sampling error must also be controlled to preserve the stability guarantees used in the IPM analysis. Since each block has constant dimension, the corresponding matrix operations can be performed efficiently whenever its approximation is refreshed.

Block Lewis-weight maintenance. We extend the dynamic leverage-score maintenance framework of [VDBLL+21] to block-diagonal scalings. The main challenge is to detect significant changes in block leverage scores under matrix-valued updates. We develop blockwise detection and verification procedures based on randomized sketches, dyadic decomposition, and lazy updates. Since each block is associated with a single scalar leverage score, the existing iterative framework for maintaining regularized Lewis weights carries over once the corresponding blockwise leverage-score approximation lemma is established.

Other block data structures. The remaining data structures largely follow existing fast IPM implementations, with modifications to accommodate blockwise updates. These include heavy-hitter detection, dual and gradient maintenance, and stochastic sampling, all implemented using blockwise versions of the existing lazy-update and bucketing techniques. We also extend inverse maintenance for the Newton systems to full positive definite block updates. Since each block has constant

dimension, such updates induce only constant-rank changes in the system matrix, allowing us to exploit block leverage-score overestimates in the maintenance procedure. Together with finite-sum smoothing and block Lewis-weight maintenance, these structures yield an efficient implementation of the idealized IPM.

3 Preliminaries

We introduce the formal notations, definitions and properties we use throughout the paper.

3.1 Notations

For a positive integer n, let $[n] = \{1, \ldots, n\}$. We use bold uppercase letters such as $\mathbf{A}$ to denote matrices, lowercase letters such as v to denote vectors. Let $\mathbf{I}_d$ denote the $d \times d$ identity matrix, $\mathbf{1}$ the all-ones vector, and e_i the i-th standard basis vector. For a vector v, we use $\|v\|_p$ for the vector ℓ_p-norm, and $\|v\|_\infty$ for the vector infinity norm. For a matrix $\mathbf{A}$, let $\mathrm{sym}(\mathbf{A})$ denote $(\mathbf{A} + \mathbf{A}^\top)/2$, and $\mathrm{asym}(\mathbf{A})$ denote $(\mathbf{A} - \mathbf{A}^\top)/2$. Also, we use $\|\mathbf{A}\|_{\mathrm{op}}$ for the operator norm, $\|\mathbf{A}\|_{\mathrm{F}}$ for the Frobenius norm, and $\|\mathbf{A}\|_\infty$ for the induced infinity norm. For symmetric matrix $\mathbf{A}$ and positive definite $\mathbf{B} \succ 0$, let $\|v\|_{\mathbf{B}} = \sqrt{v^\top \mathbf{B} v}$, $\|\mathbf{A}\|_{\mathbf{B}^{-1}} = \left\|\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}}\right\|_{\mathrm{op}}$. For a vector v, let $\mathsf{diag}(v)$ denote the diagonal matrix with diagonal entries given by v, and let $\|\cdot\|_v$ denote $\|\cdot\|_{\mathsf{diag}(v)}$.

Block vectors and matrices. Let $\mathrm{G} = (G_1, \ldots, G_m)$ be the block partition of $[n]$, where $|G_i| = n_i$ and $\sum_{i=1}^m n_i = n$. For a vector $v \in \mathbb{R}^n$, let $v_i \in \mathbb{R}^{n_i}$ denote its i-th block. Let $\|v\|_{2,*} = \left\|(\|v_i\|_2)_{i\in[m]}\right\|_*$ denote its block norm, where $*$ specifies the norm type. Similarly, for a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, let $\mathbf{A}_i \in \mathbb{R}^{n_i \times d}$ denote its i-th row block. For a vector $w \in \mathbb{R}^m$, define

$$\mathsf{diag}_{\mathrm{G}}(w) = \begin{bmatrix} w_1\mathbf{I}_{n_1} & & \\ & \ddots & \\ & & w_m\mathbf{I}_{n_m} \end{bmatrix}.$$

We use the bold notation $\mathbf{W}$ for both $\mathsf{diag}(w)$ and $\mathsf{diag}_{\mathrm{G}}(w)$ when it can be inferred from the context. For a matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, let $\mathbf{M}_{ij} \in \mathbb{R}^{n_i \times n_j}$ denote its (i,j)-th block and $\mathbf{M}_i = \mathbf{M}_{ii}$ its i-th principal block. We define the block trace $\mathrm{tr}_{\mathrm{G}}(\mathbf{M}) = (\mathrm{tr}(\mathbf{M}_1), \ldots, \mathrm{tr}(\mathbf{M}_m)) \in \mathbb{R}^m$. For two matrices $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{n\times n}$, we define their blockwise Frobenius product $\mathbf{M} \circ_{\mathrm{G}} \mathbf{N} \in \mathbb{R}^{m\times m}$ by $(\mathbf{M} \circ_{\mathrm{G}} \mathbf{N})_{ij} = \mathrm{tr}(\mathbf{M}_{ij}\mathbf{N}_{ij}^\top)$. Also, we use $\mathbf{E}_i = \mathsf{diag}_{\mathrm{G}}(e_i)$.

Projection Matrix and Leverage Score. For a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ with full column rank, let $\mathbf{P}(\mathbf{A}) = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$ denote the orthogonal projection matrix onto its column space. The leverage score of j-th row is simply $\|e_j^\top \mathbf{P}(\mathbf{A})\|_2^2$. We define its block leverage score vector as $\sigma(\mathbf{A}) = \mathrm{tr}_{\mathrm{G}}(\mathbf{P}(\mathbf{A})) \in \mathbb{R}^m$.

Approximation and comparison. For two vectors a, b of the same dimension, we write $a \approx_\epsilon b$ if $e^{-\epsilon}b_i \le a_i \le e^\epsilon b_i$ for any i. For two matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \succeq \mathbf{B}$ if and only if $\mathbf{A} - \mathbf{B}$ is PSD. We write $\mathbf{A} \approx_\epsilon \mathbf{B}$ if $e^{-\epsilon}\mathbf{B} \preceq \mathbf{A} \preceq e^\epsilon\mathbf{B}$. Let $C \ge 10000$ be a sufficiently large universal constant, larger than all absolute constants hidden in $\lesssim$, $O(\cdot)$, and $\widetilde{O}(\cdot)$ throughout the paper. Here $\widetilde{O}(\cdot)$ additionally hides all polylogarithmic factors. Although the block sizes $n_i = O(1)$ can be absorbed

into the final running-time bound, we explicitly track their dependence in intermediate accuracy estimates and verify that it is absorbed by our parameter choices.

High probability event. We say an event holds with high probability or simply w.h.p. if it happens with probability at least $1 - m^{-10}$. As a consequence, we ensure still ut holds after all possible union bound or other probability computations throughout the paper.

3.2 Self-concordant Barriers

We first recall the definition and several standard properties of self-concordant barriers.

Definition 2 (Self-concordance)**.** Let $K \subseteq \mathbb{R}^n$ be a convex set, and $\phi : K \to \mathbb{R}$ be a convex function. ϕ is ν-self-concordant if

1. For all $x \in \operatorname{int} K$ and $h \in \mathbb{R}^n$, it holds that $\left\|\nabla^3 \phi(x)[h,h,h]\right\| \le 2\|h\|^3_{\nabla^2\phi(x)}$;
2. For all $x \in \operatorname{int} K$, $\nabla\phi(x)^\top \nabla^2\phi(x)^{-1}\nabla\phi(x) \le \nu$.

Lemma 3. *Let ϕ be a ν-self-concordant barrier on K. For any $x \in \operatorname{int} K$ and $h \in \mathbb{R}^n$, let $r = \|h\|_{\nabla^2\phi(x)} < 1$. Then the following properties hold:*

1. *Local feasibility: $x + h \in K$;*
2. *Hessian comparison: $(1-r)^2\nabla^2\phi(x) \preceq \nabla^2\phi(x+h) \preceq (1-r)^{-2}\nabla^2\phi(x)$;*
3. *First-order smoothness: $\|\nabla\phi(x+h) - \nabla\phi(x)\|_{\nabla^2\phi(x)^{-1}} \le \frac{r}{1-r}$;*
4. *Second-order error: $\left\|\nabla\phi(x+h) - \nabla\phi(x) - \nabla^2\phi(x)h\right\|_{\nabla^2\phi(x)^{-1}} \le \frac{r^2}{1-r}$.*

3.3 Fenchel Conjugate and Bregman Divergence

We next recall the Fenchel conjugate and its properties.

Definition 4. Let $\phi : K \to \mathbb{R}$ be a convex function. Its Fenchel conjugate ϕ^* evaluated at y is defined as $\phi^*(y) = \sup_{x\in K} \langle x, y\rangle - \phi(x)$. If the maximizer exists and is unique, denote $x(y) = \operatorname{argmax}_{x\in K} \langle x, y\rangle - \phi(x)$.

The following identities and local estimates are standard consequences of Fenchel conjugate of a self-concordance function.

Lemma 5. *Let ϕ be a ν-self-concordant barrier on K and ϕ^* be its Fenchel conjugate on K^*. Then ϕ^* is also self-concordant. The following facts hold for $x \in K$ and $y \in K^*$*

$$x(y) = (\nabla\phi)^{-1}(y), \qquad \nabla\phi^*(y) = x(y), \qquad \nabla^2\phi^*(y) = \nabla^2\phi(x(y))^{-1}.$$

Lemma 6. *Let ϕ be a self-concordant barrier on K. For any points $a \in K^*$ and $h \in \mathbb{R}^n$, let $r = \|h\|_{\nabla^2\phi^*(a)} < 1$. Then the following property holds:*

$$\|\nabla\phi^*(a+h) - \nabla\phi^*(a)\|_{\nabla^2\phi^*(a)^{-1}} \le \frac{r}{1-r},$$

Moreover,

$$\left\|\nabla\phi^*(a+h) - \nabla\phi^*(a) - \nabla^2\phi^*(a)h\right\|_{\nabla^2\phi^*(a)^{-1}} \le \frac{r^2}{1-r}.$$

We also recall the definition of a Bregman divergence and present an estimate when the potential function is a Fenchel conjugate of a self-concordant barrier. The estimate gives a local quadratic approximation of the Bregman divergence.

Definition 7 (Bregman divergence)**.** Let $f : K \to \mathbb{R}$ be a differentiable convex function. Its Bregman divergence between two points $x, y \in K$ is defined as $D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle$.

Lemma 8. *Let ϕ be a self-concordant function and ϕ^* be its Fenchel conjugate. For any $a, b \in \operatorname{dom} \phi^*$, let $r = \|a - b\|_{\nabla^2 \phi^*(b)} \le 1/2$. Then the following property holds for the Bregman divergence*

$$\frac{1}{2}r^2 - r^3 \le D_{\phi^*}(a, b) \le \frac{1}{2}r^2 + r^3.$$

3.4 Randomized smoothing

Finally, we introduce the uniform randomized smoothing with a local Hessian preconditioning and a lemma to characterize the major properties. The lemma is a simple adoption of the standard result from [DBW12].

Definition 9 (Randomized smoothing)**.** Let $\mathbf{C} : K \subseteq \mathbb{R}^d \to \mathbb{R}^{n \times n}$ be a symmetric and PSD matrix-value function and $\mathbf{M} \succ 0$ be a local symmetric reference matrix. For any x, if $x + \sigma \mathbf{M}^{-\frac{1}{2}} B_d \subseteq K$ holds, the local uniform smoothing at x and smoothing scale $\sigma > 0$ is defined as

$$\mathbf{C}_{\sigma,\mathbf{M}}(x) := \mathbb{E}_{u \sim \mathrm{Unif}(B_d)} \left[\mathbf{C}(x + \sigma \mathbf{M}^{-\frac{1}{2}} u) \right]$$

where B_d is the centered unit ball in $\mathbb{R}^d$.

The following lemma presents the standard approximation and smoothness properties of local uniform smoothing.

Lemma 10. *Let $\mathbf{C} : K \subseteq \mathbb{R}^d \to \mathbb{R}^{n \times n}$ be a matrix-valued function and $\mathbf{M} \succ 0$ be a local reference matrix associated with x. Suppose that, for any y, $z \in x + (r + \sigma)\mathbf{M}^{-\frac{1}{2}} B_d$,*

$$\|\mathbf{C}(y) - \mathbf{C}(z)\|_{\mathbf{C}(x)^{-1}} \le G \|y - z\|_{\mathbf{M}}.$$

Then following properties hold for its local uniform smoothing $\mathbf{C}_{\sigma,\mathbf{M}}(\cdot)$ and any y, $z \in x + r\mathbf{M}^{-\frac{1}{2}} B_d$:

1. *$\mathbf{C}_{\sigma,\mathbf{M}}(\cdot)$ is differentiable with $\mathrm{D}\mathbf{C}_{\sigma,\mathbf{M}}(y)[h] = \frac{d}{\sigma} \mathbb{E}_{u \sim \mathrm{Unif}(\partial B_d)} \left[\mathbf{C}(y + \sigma \mathbf{M}^{-\frac{1}{2}} u) \left\langle u, \mathbf{M}^{\frac{1}{2}} h \right\rangle \right]$;*
2. *$\|\mathbf{C}_{\sigma,\mathbf{M}}(z) - \mathbf{C}_{\sigma,\mathbf{M}}(y) - \mathrm{D}\mathbf{C}_{\sigma,\mathbf{M}}(y)[z - y]\|_{\mathbf{C}(x)^{-1}} \le \frac{G\sqrt{d}}{2\sigma} \|y - z\|_{\mathbf{M}}^2$,*

4 Block Lewis-weight IPM

Here we develop a robust short-step IPM for the product-domain optimization problem (2). Before proceeding to the algorithm, we introduce several necessary concepts and definitions.

Block Lewis weights and smoothing scaling. We use regularized block Lewis weights to define the weighted central path. In particular, a smoothed Hessian-based preconditioner is employed to control the stability of Lewis weights under randomized sparse primal updates, without requiring higher-order self-concordance of the barrier functions.

Definition 11 (Smoothed preconditioner)**.** Let x be a point in K and $\mathbf{M} \in \mathbb{R}^{n\times n}$ be a symmetric block-diagonal matrix with $\mathbf{M} \succ 0$. Let $\sigma > 0$ be the smoothing scale. The smoothed preconditioner for the i-th block is defined as

$$\widetilde{\mathbf{C}}_i(x_i, \mathbf{M}_i) = \widetilde{\mathbf{S}}_i(x_i, \mathbf{M}_i)^{\frac{1}{2}}, \qquad \text{where} \qquad \widetilde{\mathbf{S}}_i(x_i, \mathbf{M}_i) := \mathbb{E}_{u\sim \mathrm{Unif}(B_{n_i})}\left[\nabla^2\phi_i\left(x_i + \sigma \mathbf{M}_i^{-\frac{1}{2}}u\right)^{-1}\right].$$

Correspondingly, we write the full matrix as $\widetilde{\mathbf{S}}(x, \mathbf{M}) := \mathsf{diag}_{\mathrm{G}}\big(\widetilde{\mathbf{S}}_1(x_1, \mathbf{M}_1), \ldots, \widetilde{\mathbf{S}}_m(x_m, \mathbf{M}_m)\big)$, and $\widetilde{\mathbf{C}}(x, \mathbf{M}) = \widetilde{\mathbf{S}}(x, \mathbf{M})^{\frac{1}{2}}$.

Definition 12 (Regularized block Lewis weight)**.** For any $p > 0$, the p-th order regularized block Lewis weight $w_p(\mathbf{A}) \in \mathbb{R}^m$ of matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ is the solution of the following system

$$w = \mathrm{tr}_{\mathrm{G}}\left(\mathbf{P}(\mathbf{W}^{\frac{1}{2}-\frac{1}{p}}\mathbf{A})\right) + v,$$

where $\mathbf{W} = \mathsf{diag}_{\mathrm{G}}(w)$, $v \in \mathbb{R}^m_{>0}$ is the regularization weight.

Definition 13. For any block-diagonal $\mathbf{C} \in \mathbb{R}^{n\times n}$, define the $\mathbf{C}$-preconditioned regularized block Lewis weight with regularization $v = \frac{d}{m}\mathbf{1}_m$ as $w(\mathbf{C}) := w_p(\mathbf{CA})$ when there is no ambiguity. For any $x \in K$, define $\tau(x)$ as $\tau(x) = w\left(\widetilde{\mathbf{C}}(x, \nabla^2\Phi(x))\right)$.

Block mixed norm. Following [vdBLSS20, VDBLL+21], we extend the mixed norms to a block version. This allows us to simultaneously control the weighted Euclidean and maximum block errors. Their corresponding dual vectors are used to construct the update direction.

Definition 14 (Mixed norm)**.** For a vector $u \in \mathbb{R}^m$, let $\|u\|_{\tau+\infty} := \|u\|_\infty + C_{\mathrm{norm}}\|u\|_\tau$ for some constant C_{norm} and

$$u^{\flat(\tau)} := \mathrm{argmax}_{\|h\|_{\tau+\infty}\le 1} h^\top u.$$

The corresponding dual norm is defined as $\|u\|^*_{\tau+\infty} := u^\top u^{\flat(\tau)}$. Moreover, for a block vector $v \in \mathbb{R}^n$, its mixed block norm is $\|v\|_{2,\tau+\infty} := \|v\|_{2,\infty} + C_{\mathrm{norm}}\|v\|_{2,\tau}$, where the block-$\tau$ and block-∞ norms are defined as

$$\|v\|_{2,\tau} := \big\|\big(\|v_i\|_2\big)_{i\in[m]}\big\|_\tau, \qquad \|v\|_{2,\infty} := \max_{i\in[m]}\|v_i\|_2.$$

The corresponding dual vector and dual norm, denoted by $v^{\flat(\tau)}$ and $\|v\|^*_{2,\tau+\infty}$, are defined in the same way.

Bregman metric and potential. We measure deviations from the weighted central path using the Bregman divergence. This choice avoids directly differentiating the varying inverse-Hessian metric and allows us to analyze residual changes using the standard self-concordance. We then use the exponential function to build the potential function.

Definition 15 (Central path residual and metric)**.** Let the vector $\mu \in \mathbb{R}^n$ be the residual of the perturbed weighted central path at the triplet (x, s, ρ), defined as

$$\mu_i(x, s, \rho) = \frac{s_i}{\rho\tau(x)_i} + \nabla\phi_i(x_i).$$

Moreover, let $\Gamma \in \mathbb{R}^m_{\geq 0}$ be the residual measure

$$\Gamma_i(\mu_i, x_i) = 2D_{\phi_i^*}(\nabla\phi_i(x_i) - \mu_i, \nabla\phi_i(x_i)).$$

Definition 16 (Potential function)**.** The centrality potential Ψ evaluated at μ and x is defined as

$$\Psi(\mu, x) = \sum_{i=1}^{m} \psi(\Gamma_i(\mu_i, x_i))$$

where $\psi(a) = e^{\lambda a}$, and $\lambda > 0$ is a penalty factor. Moreover, the normalized gradient with respect to μ is defined as $\nabla^{\text{loc}}_{\mu}\Psi(\mu, x) = \nabla^2\Phi(x)^{\frac{1}{2}}\nabla_{\mu}\Psi(\mu, x)$,

Potential function-based analysis indicates that the triplet (x, s, ρ) obeys the following condition throughout the algorithm. In particular, the first condition ensures that the trajectory is close to the exact central path. It should also be noticed that only an approximate primal feasibility is guaranteed. A post-processing technique will restore the exact feasibility. We postpone the discussion to Section 6.

Definition 17. For any $\epsilon \in (0, 1/100]$, the triplet (x, s, ρ) is ϵ-centered if the following properties hold

1. Approximate centrality: $\Gamma_i(\mu_i, x_i) \leq \epsilon^2$ for $\mu = \mu(x, s, \rho)$ and any $i \in [m]$;
2. Approximate primal feasibility: $\left\|\mathbf{A}^\top x - b\right\|_{(\mathbf{A}^\top\mathbf{T}^{-1}\nabla^2\Phi(x)^{-1}\mathbf{A})^{-1}} \leq \frac{\alpha\beta}{2C_{\text{norm}}}$,;
3. Dual feasibility: $\mathbf{A}y + s = c$ for some $y \in \mathbb{R}^d$.

Moreover, a robust IPM allows the algorithm to use approximations for ideal quantities. The following definition specifies the approximation guarantees required by our analysis.

Definition 18. The pair $(\overline{x}, \overline{\tau})$ is an α-approximation of the ideal (x, τ) if the following conditions hold:

$$\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}(x_i - \overline{x}_i)\right\|_2 \leq \alpha, \qquad \left\|\mathbf{T}(x)^{-1}(\tau(x) - \overline{\tau})\right\|_\infty \leq \alpha.$$

Similarly, $\overline{\mu}$ is a 10α-approximation of μ if

$$\left\|\nabla^2\phi_i(x_i)^{-\frac{1}{2}}(\mu_i - \overline{\mu}_i)\right\|_2 \leq 10\alpha.$$

Block sampling. The projection correction in the primal update is dense and expensive to maintain explicitly. Similar to [vdBLN+20,VDBLL+21,LPT+25], we randomly subsample this correction at the block level to reduce the update cost.

Definition 19. Let $C_{\text{Var}} \geq C_{\text{norm}}$ be a sufficiently large constant and $\delta \in \mathbb{R}^n$ be a vector. $\mathbf{R} = \mathsf{diag}_{\text{G}}(R) \in \mathbb{R}^{n \times n}$ is a valid sampling block-diagonal matrix associated with δ, if the following properties hold:

1. Expectation: $\mathbb{E}[\mathbf{R}] = \mathbf{I}_n$;
2. Variance: $\text{Var}\left[R_i \left\|\delta_i\right\|_2\right] \leq \beta^2 \sqrt{\frac{d}{m}} \frac{\left\|\delta_i\right\|_2 \left\|\delta\right\|_2}{C_{\text{Var}}^2}$, $\text{Var}\left[R_i\right] \leq 2\tau_i^{-1}$;
3. Covariance: $\mathbb{E}[R_i R_j] \leq 2$ for any $i \neq j$;
4. Maximum: w.h.p. $\left\|\mathbf{R}\delta - \delta\right\|_{2,\infty} \leq \beta \sqrt{\frac{d}{m}} \frac{\left\|\delta\right\|_2}{C_{\text{Var}}^2}$;
5. Spectral Approximation: w.h.p.

$$\mathbf{A}^\top \nabla^2 \Phi(x)^{-\frac{1}{2}} \mathbf{T}^{-\frac{1}{2}} \mathbf{R} \mathbf{T}^{-\frac{1}{2}} \nabla^2 \Phi(x)^{-\frac{1}{2}} \mathbf{A} \approx_\alpha \mathbf{A}^\top \mathbf{T}^{-1} \nabla^2 \Phi(x)^{-1} \mathbf{A}.$$

Note that α, β are certain constants required by the ideal algorithm.

Short-step Algorithm. We now present the ideal ShortStep algorithm, which describes a single iteration of our block path-following method. The complete path-following algorithm will be developed as an efficient implementation of this ideal procedure in Section 6.

Our convergence analysis proceeds by induction over successive calls to ShortStep. Assuming that the current iterate is ϵ-centered, the key step is to establish a contraction of the expected potential under one iteration and the prescribed reduction of the path parameter, as stated in the following lemma.

Lemma 20. *Suppose that (x, s, ρ) is ϵ-centered as in Definition 17. Let $(x^{(\text{new})}, s^{(\text{new})})$ be the output of Algorithm 1, and $\rho^{(\text{new})} = (1 - r)\rho$ with $r = \frac{\alpha\beta}{4C_{norm}\sqrt{d}}$. Then it holds that*

$$\mathbb{E}[\Psi(\mu^{(\text{new})}, x^{(\text{new})})] \leq \left(1 - \frac{\theta^2}{20CC_{norm}C_\lambda\sqrt{d}}\right) \Psi(\mu, x) + m$$

where $\mu = \mu(x, s, \rho)$ and $\mu^{(\text{new})} = \mu(x^{(\text{new})}, s^{(\text{new})}, \rho^{(\text{new})})$ are defined in (7).

We must show that each call to ShortStep preserves the ϵ-centrality conditions. Approximate Centrality will follow readily from Lemma 20, as an application of the Markov inequality ensures $\Psi(\mu, x) \leq \text{poly}(m)$, and hence $\left\|\Gamma\right\|_\infty \leq \epsilon^2$ bye our parameter choices. The Dual Feasibility follows immediately from the update rule. It remains to establish Approximate Primal Feasibility and construct a valid block sampling scheme. We defer these to Section 4.5.

Algorithm 1: Short-Step

parameters

Set $\theta = \frac{1}{10\log(m)}$, $p = 1-\theta$, $\epsilon = \frac{\theta}{C}$, $C_{\text{norm}} = \frac{C}{\theta}$, $C_\lambda = 30\log(10m)$, $\lambda = \frac{C_\lambda}{\epsilon^2}$, $\beta = \frac{\theta}{CC_\lambda\sqrt{\lambda}}$, $\alpha = \frac{\beta}{C}$, $\sigma = \theta/C$.

procedure *ShortStep*$(x, s, \rho, \rho^{(\text{new})})$

Assert $\left|\rho^{(\text{new})} - \rho\right| \le r\rho$.

Let $(\overline{x}, \overline{\tau})$ be an α-approximation of $(x, \tau(x))$ per Definition 18 .

Let $\mu = \mu(x, s, \rho)$.

Let $\overline{\mu}$ be a 10α-approximation of μ.

Let $g = -\beta\nabla_\mu^{\text{loc}}\Psi(\overline{\mu}, \overline{x})^{\flat(\overline{\tau})}$.

Let $\mathbf{H} = \mathbf{A}^\top\overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-1}\mathbf{A}$, and $\overline{\mathbf{H}} \approx_\alpha \mathbf{H}$.

Let $\delta_1 = \overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g$.

Let $\delta_2 = \overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}(\mathbf{A}^\top x - b)$

Let $\delta_r = \delta_1 + \delta_2$.

Let $\mathbf{R} \in \mathbb{R}^{n\times n}$ be some random block diagonal-matrix, per Definition 19.

$\delta_x = \nabla^2\Phi(\overline{x})^{-\frac{1}{2}}(g - \mathbf{R}\delta_r)$.

$\delta_s = \rho\overline{\mathbf{T}}\nabla^2\Phi(\overline{x})^{\frac{1}{2}}\delta_1$.

$(x^{(\text{new})}, s^{(\text{new})}) \leftarrow (x + \delta_x, s + \delta_s)$.

return $(x^{(\text{new})}, s^{(\text{new})})$

Outline of Analysis. The following subsections are devoted to the proof of Lemma 20. We begin with important auxiliary lemmas in Section 4.1. The frequently-referenced results lay down the foundation for our randomized smoothing and Lewis-weight stability analyses. Section 4.2 then establishes stability bounds for the variables x, s, $\mathbf{C}$, τ, and μ. In particular, we defer several proofs to Appendix B. The deferred lemmas and proofs range from straightforward block extensions of their scalar counterparts in [VDBLL+21], to more technically involved adaptations. This allows us to focus on our major technical innovations when presenting the main body. The stability and approximation bounds for the Bregman metric are developed in Section 4.3. Finally, we prove Lemma 20 through a novel Bregman-based potential analysis in Section 4.4.

4.1 Supporting and Technical Lemma

In this subsection, we provide some basic tools for the analysis our IPM. The first helper lemma indicates that the Lewis weights are close when the square of their corresponding preconditioning scaling are blockwise close.

Lemma 21. *Let $p \in (0, 2]$. For any two symmetric and block-diagonal SPD matrices $\mathbf{C}, \mathbf{C}' \in \mathbb{R}^{n\times n}$, if $(\mathbf{C}_i)^2 \approx_{2\epsilon} (\mathbf{C}'_i)^2$ is true for some $\epsilon > 0$ and any block $i \in [m]$, then $w \approx_{8\epsilon} w'$, where $w = w(\mathbf{C})$ and $w' = w(\mathbf{C}')$.*

Proof. In this proof, $\mathbf{W}$ will always denote the block-identity matrices of w. Define mapping $f(w, \mathbf{C})$ as

$$(f(w,\mathbf{C}))_i = \left(\operatorname{tr}\left(\mathbf{C}_i\mathbf{A}_i\left(\mathbf{A}^\top\mathbf{C}\mathbf{W}^{1-\frac{2}{p}}\mathbf{C}\mathbf{A}\right)^{-1}\mathbf{A}_i^\top\mathbf{C}_i\right) + v_i w_i^{2/p-1}\right)^{p/2}$$

for any i, and consider a sequence of weights $w_k \in \mathbb{R}^m$ generated by iteration $w_{k+1} = f(w_k, \mathbf{C}')$ with initialization $w_0 = w$. Our aim is to show $w' = \lim_{k\to\infty} w_k \approx_{\epsilon'} w_0 = w$, for some ϵ', as it is easy to see w and w' are the fixed point of $f(\cdot, \mathbf{C})$ and $f(\cdot, \mathbf{C}')$.

The base case requires to prove $w_1 = f(w_0, \mathbf{C}') \approx f(w_0, \mathbf{C}) = w_0$. To see this, the following holds from square approximation $(\mathbf{C}_i)^2 \approx_{2\epsilon} (\mathbf{C}'_i)^2$

$$(\mathbf{C}\mathbf{W}^{1-\frac{2}{p}}\mathbf{C})_i = w_i^{1-2/p}(\mathbf{C}_i)^2 \approx_{2\epsilon} w_i^{1-2/p}(\mathbf{C}'_i)^2 = (\mathbf{C}'\mathbf{W}^{1-\frac{2}{p}}\mathbf{C}')_i.$$

This implies $\mathbf{A}^\top\mathbf{C}\mathbf{W}^{1-\frac{2}{p}}\mathbf{C}\mathbf{A} \approx_{2\epsilon} \mathbf{A}^\top\mathbf{C}'\mathbf{W}^{1-\frac{2}{p}}\mathbf{C}'\mathbf{A}$ due to the fact that $\mathbf{A}^\top\mathbf{B}\mathbf{A} \approx \mathbf{A}^\top\mathbf{B}'\mathbf{A}$ holds if $\mathbf{B} \approx \mathbf{B}'$ and both $\mathbf{B}, \mathbf{B}' \succeq 0$. We compute

$$\begin{aligned}
(f(w, \mathbf{C}))_i &= \left(\left\langle \left(\mathbf{A}^\top\mathbf{C}\mathbf{W}^{1-\frac{2}{p}}\mathbf{C}\mathbf{A}\right)^{-1}, \mathbf{A}_i^\top\mathbf{C}_i^2\mathbf{A}_i \right\rangle + v_i w_i^{2/p-1}\right)^{p/2} \\
&\approx_{2p\epsilon} \left(\left\langle \left(\mathbf{A}^\top\mathbf{C}'\mathbf{W}^{1-\frac{2}{p}}\mathbf{C}'\mathbf{A}\right)^{-1}, \mathbf{A}_i^\top(\mathbf{C}_i)^2\mathbf{A}_i \right\rangle + v_i w_i^{2/p-1}\right)^{p/2} \\
&\approx_{4p\epsilon} \left(\left\langle \left(\mathbf{A}^\top\mathbf{C}'\mathbf{W}^{1-\frac{2}{p}}\mathbf{C}'\mathbf{A}\right)^{-1}, \mathbf{A}_i^\top(\mathbf{C}'_i)^2\mathbf{A}_i \right\rangle + v_i w_i^{2/p-1}\right)^{p/2} \\
&= (f(w, \mathbf{C}'))_i.
\end{aligned}$$

where we use the square approximation $(\mathbf{C}_i)^2 \approx_{2\epsilon} (\mathbf{C}'_i)^2$ again in the second line.

Now, suppose $w_k \approx_\delta w_{k-1}$ for some certain $\delta > 0$. We aim to find the relation between w_{k+1} and w_k. From $w_k \approx_\delta w_{k-1}$, it immediately follows that $\mathbf{W}_k^{1-\frac{2}{p}} \approx_{|1-\frac{2}{p}|\delta} \mathbf{W}_{k-1}^{1-\frac{2}{p}}$. We continue as

$$\mathrm{tr}_{\mathrm{G}}\left(\mathbf{C}'\mathbf{A}\left(\mathbf{A}^\top\mathbf{C}'\mathbf{W}_k^{1-\frac{2}{p}}\mathbf{C}'\mathbf{A}\right)^{-1}\mathbf{A}^\top\mathbf{C}'\right) \approx_{|1-2/p|\delta} \mathrm{tr}_{\mathrm{G}}\left(\mathbf{C}'\mathbf{A}\left(\mathbf{A}^\top\mathbf{C}'\mathbf{W}_{k-1}^{1-\frac{2}{p}}\mathbf{C}'\mathbf{A}\right)^{-1}\mathbf{A}^\top\mathbf{C}'\right)$$

It follows that by considering the regularization and constant factor

$$w_{k+1} = f(w_k, \mathbf{C}') \approx_{|(1-p/2)|\delta} f(w_{k-1}, \mathbf{C}') = w_k.$$

This suggests, with $p \in (0, 2)$, the approximation ratio between w_{k+1} and w_k shrinks by a rate of $|1 - p/2|$ when the iteration k grows. As a consequence, $w = w_0 \approx_{\epsilon'} \lim_{k\to\infty} w_k = w'$ with

$$\epsilon' = 4p\epsilon \cdot \lim_{k\to\infty} \sum_k |1 - p/2|^k = 4p\epsilon \cdot \lim_{k\to\infty} \frac{1 - (1 - p/2)^{k+1}}{1 - (1 - p/2)} = 8\epsilon.$$

when $p \in (0, 2]$. □

The next lemma provides the norm bound for asymetric matrices of the form $\mathbf{C}^{-1}\mathbf{B}$.

Lemma 22. *Let $\mathbf{A} = \mathbf{C}\mathbf{B}^{-1}$, where $\mathbf{B} \succ 0$ and $\mathbf{C}$ are symmetric matrices of dimension d. Then it holds that $\|\mathbf{A}\|_{\mathrm{op}}^2 \le 2d\,\|\mathrm{sym}(\mathbf{A})\|_{\mathrm{op}}^2$.*

Proof. Since $\mathbf{A}$ is similar to the symmetric matrix $\mathbf{B}^{-\frac{1}{2}}\mathbf{C}\mathbf{B}^{-\frac{1}{2}}$, we have

$$\left\|\mathrm{sym}(\mathbf{A})\right\|_{\mathrm{F}}^2 - \left\|\mathrm{asym}(\mathbf{A})\right\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{A}^2) = \left\|\mathbf{B}^{-\frac{1}{2}}\mathbf{C}\mathbf{B}^{-\frac{1}{2}}\right\|_{\mathrm{F}}^2.$$

As a result, it holds that

$$\|\mathbf{A}\|_{\mathrm{op}}^2 \le \|\mathbf{A}\|_{\mathrm{F}}^2 = \|\mathrm{sym}(\mathbf{A})\|_{\mathrm{F}}^2 + \|\mathrm{asym}(\mathbf{A})\|_{\mathrm{F}}^2 \le 2\,\|\mathrm{sym}(\mathbf{A})\|_{\mathrm{F}}^2 \le 2d\,\|\mathrm{sym}(\mathbf{A})\|_{\mathrm{op}}^2$$

and concludes the proof. □

We now state two important auxiliary lemmas. They characterize the properties of the ideal randomized smoothing in Definition 9 when we vary the variable and the referencing local metric, respectively.

Lemma 23. *Let $\phi : K \subseteq \mathbb{R}^d \to \mathbb{R}$ be a 1-self-concordant function. Let x be a fixed reference point, $\mathbf{M} = \nabla^2\phi(x)$ be a reference matrix, and $\sigma > 0$ be the smoothing scale. For any y, define $\mathbf{S}(y) = \nabla^2\phi(y)^{-1}$, $\mathbf{C}(y) = \nabla^2\phi(y)^{-\frac{1}{2}}$ and*

$$\widetilde{\mathbf{C}}(y, \mathbf{M}) = \widetilde{\mathbf{S}}(y, \mathbf{M})^{\frac{1}{2}}, \qquad \text{where} \qquad \widetilde{\mathbf{S}}(y, \mathbf{M}) = \mathbb{E}_{u\sim \mathrm{Unif}(B_d)} \left[\mathbf{S}\left(y + \sigma \mathbf{M}^{-\frac{1}{2}} u\right)\right].$$

If $r_{\max} < 1/4$, $\sigma < 1/4$, then $\widetilde{\mathbf{C}}(y, \mathbf{M})$ is well defined for every y with $r := \|x - y\|_{\mathbf{M}} \le r_{\max}$, and the following properties hold for reference x and any such y

1. $\widetilde{\mathbf{S}}(x, \mathbf{M}) \approx_{4\sigma} \mathbf{S}(x)$, $\widetilde{\mathbf{C}}(x, \mathbf{M}) \approx_{2\sigma} \mathbf{C}(x)$;
2. $\widetilde{\mathbf{S}}(y, \mathbf{M}) \approx_{4r_{\max}} \widetilde{\mathbf{S}}(x, \mathbf{M})$, $\widetilde{\mathbf{C}}(y, \mathbf{M}) \approx_{2r_{\max}} \widetilde{\mathbf{C}}(x, \mathbf{M})$;
3. $\left\|\mathrm{sym}\big(\widetilde{\mathbf{C}}(y, \mathbf{M})^{-1} \mathrm{D}\widetilde{\mathbf{C}}(y, \mathbf{M})[h]\big)\right\|_{\mathrm{op}} \le (1 + 2(r_{\max} + \sigma)) \|h\|_{\mathbf{M}}$;
4. $\left\|\mathrm{sym}\big(\widetilde{\mathbf{C}}(x, \mathbf{M})^{-1}\big(\widetilde{\mathbf{C}}(y, \mathbf{M}) - \widetilde{\mathbf{C}}(x, \mathbf{M})\big)\big)\right\|_{\mathrm{op}} \le 4r + 16dr^2$;
5. $\left\|\mathrm{sym}\big(\big(\widetilde{\mathbf{C}}(y, \mathbf{M}) - \widetilde{\mathbf{C}}(x, \mathbf{M}) - \mathrm{D}\widetilde{\mathbf{C}}(x, \mathbf{M})[y - x]\big)\widetilde{\mathbf{C}}(x, \mathbf{M})^{-1}\big)\right\|_{\mathrm{op}} \le \left(8d + \frac{20\sqrt{d}}{\sigma}\right) r^2$.

Proof. We first verify that the smoothed matrices are well-defined at such an y. As $\|y - x\|_{\mathbf{M}} \le r_{\max} < 1/4$, we assert $y \in K$ using Point 1 in Lemma 3. Also, for any $u \in B_d$, it holds that

$$\left\|y + \sigma \mathbf{M}^{-\frac{1}{2}} u - x\right\|_{\mathbf{M}} \le \|y - x\|_{\mathbf{M}} + \sigma \|u\|_2 \le r_{\max} + \sigma < 1/2$$

The same lemma also indicates $y + \sigma \mathbf{M}^{-\frac{1}{2}} u \in K$. Therefore both $\widetilde{\mathbf{S}}(y, \mathbf{M})$ and $\widetilde{\mathbf{C}}(y, \mathbf{M})$ are legitimate because $\widetilde{\mathbf{S}} \succ 0$ is clearly true from the definition.

For the first four properties, we will use intensively the self-concordance and the Hessian comparison in Point 2 in Lemma 3. We do not repeat the lemma referencing in the rest part of the proof. Fix y with $\|y - x\|_{\mathbf{M}} = r < 1/4$, Hessian comparison indicates $\nabla^2\phi(y) \preceq (1 - r)^{-2}\nabla^2\phi(x) \preceq \frac{16}{9}\mathbf{M}$. Now for any $u \in B_d$, the same argument indicates

$$\nabla^2\phi(y + \sigma \mathbf{M}^{-\frac{1}{2}} u)^{-1} \preceq \left(1 - \sigma \|\mathbf{M}^{-\frac{1}{2}} u\|_{\nabla^2\phi(y)}\right)^{-2} \nabla^2\phi(y)^{-1} \preceq e^{4\sigma}\nabla^2\phi(y)^{-1}$$

The corresponding lower bound can be obtained in a similar fashion. Taking expectation over u gives the first result $\widetilde{\mathbf{S}}(y, \mathbf{M}) \approx_{4\sigma} \mathbf{S}(y)$, and, in addition, $\widetilde{\mathbf{C}}(y, \mathbf{M}) \approx_{2\sigma} \mathbf{C}(y)$ by the monotonicity of the matrix root. To prove the second property, we use Hessian comparison again and get $\nabla^2\phi(x + \sigma \mathbf{M}^{-\frac{1}{2}} u) \preceq (1 - \sigma)^{-2}\nabla^2\phi(x) \preceq \frac{16}{9}\nabla^2\phi(x)$ as $\sigma < 1/4$. As a consequence, it holds that

$$\begin{aligned} \nabla^2\phi(y + \sigma \mathbf{M}^{-\frac{1}{2}} u)^{-1} &\preceq \left(1 - \|y - x\|_{\nabla^2\phi(x + \sigma \mathbf{M}^{-1/2} u)}\right)^{-2} \nabla^2\phi(x + \sigma \mathbf{M}^{-\frac{1}{2}} u)^{-1} \\ &\preceq (1 + 2r)^2 \nabla^2\phi(x + \sigma \mathbf{M}^{-\frac{1}{2}} u)^{-1} \preceq e^{4r}\nabla^2\phi(x + \sigma \mathbf{M}^{-\frac{1}{2}} u)^{-1} \end{aligned} \tag{10}$$

and a similar lower bound. This is immediately $\widetilde{\mathbf{S}}(y, \mathbf{M}) \approx_{4r} \widetilde{\mathbf{S}}(x, \mathbf{M})$ by taking expectation, and also $\widetilde{\mathbf{C}}(y, \mathbf{M}) \approx_{2r} \widetilde{\mathbf{C}}(x, \mathbf{M})$ by taking square root.

Now we prove the third property. Let $z = y + \sigma \mathbf{M}^{-\frac{1}{2}} u$. It holds that

$$\mathrm{D}\mathbf{S}(z)[h] \preceq 2\|h\|_{\nabla^2\phi(z)}\mathbf{S}(z) \preceq 2(1-(r+\sigma))^{-1}\|h\|_{\nabla^2\phi(z)}\mathbf{S}(x) \preceq 2(1+2(r+\sigma))\|h\|_{\mathbf{M}}\mathbf{S}(z).$$

where we use the definition of self-concordance for the first inequality, and Hessian comparison in the rest. Together with a lower bound obtained in the same way, we get

$$\left\|\mathrm{D}\widetilde{\mathbf{S}}(y)[h]\right\|_{\widetilde{\mathbf{S}}(y)^{-1}} \leq 2(1+2(r+\sigma))\left\|h\right\|_{\mathbf{M}}.$$

by taking expectation over both sides. On the other hand, since $\widetilde{\mathbf{S}} = \widetilde{\mathbf{C}}^2$, we have the following identity

$$\mathrm{D}\widetilde{\mathbf{S}}(y,\mathbf{M})[h] = \mathrm{D}\widetilde{\mathbf{C}}(y,\mathbf{M})[h]\widetilde{\mathbf{C}}(y,\mathbf{M}) + \widetilde{\mathbf{C}}(y,\mathbf{M})\mathrm{D}\widetilde{\mathbf{C}}(y,\mathbf{M})[h]$$

This suggests $\widetilde{\mathbf{C}}(y,\mathbf{M})^{-1}\mathrm{D}\widetilde{\mathbf{S}}(y,\mathbf{M})[h]\widetilde{\mathbf{C}}(y,\mathbf{M})^{-1} = 2\cdot\mathrm{sym}\big(\widetilde{\mathbf{C}}(y,\mathbf{M})^{-1}\mathrm{D}\widetilde{\mathbf{C}}(y,\mathbf{M})[h]\big)$. Consequently,

$$\left\|\mathrm{sym}\big(\widetilde{\mathbf{C}}(y,\mathbf{M})^{-1}\mathrm{D}\widetilde{\mathbf{C}}(y,\mathbf{M})[h]\big)\right\|_{\mathrm{op}} = \frac{1}{2}\left\|\mathrm{D}\widetilde{\mathbf{S}}(y,\mathbf{M})[h]\right\|_{\widetilde{\mathbf{S}}(y,\mathbf{M})^{-1}} \leq \big(1+2(r_{\max}+\sigma)\big)\left\|h\right\|_{\mathbf{M}}.$$

For the fourth property, we use $\widetilde{\mathbf{S}} = \widetilde{\mathbf{C}}^2$ again and obtain the following inequality

$$\begin{aligned}2\mathrm{sym}(\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}(\widetilde{\mathbf{C}}(y,\mathbf{M}) - \widetilde{\mathbf{C}}(x,\mathbf{M}))) &= \widetilde{\mathbf{S}}(x,\mathbf{M})^{-\frac{1}{2}}(\widetilde{\mathbf{S}}(y,\mathbf{M}) - \widetilde{\mathbf{S}}(x,\mathbf{M}))\widetilde{\mathbf{S}}(x,\mathbf{M})^{-\frac{1}{2}} \\ &\quad - \widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}(\widetilde{\mathbf{C}}(y,\mathbf{M}) - \widetilde{\mathbf{C}}(x,\mathbf{M}))^2\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}.\end{aligned}$$

In particular, the second property and $r \leq 1/4$ indicate

$$\left\|\widetilde{\mathbf{S}}(x,\mathbf{M})^{-\frac{1}{2}}(\widetilde{\mathbf{S}}(y,\mathbf{M}) - \widetilde{\mathbf{S}}(x,\mathbf{M}))\widetilde{\mathbf{S}}(x,\mathbf{M})^{-\frac{1}{2}}\right\|_{\mathrm{op}} \leq 4r + 4r^2,$$

as well as

$$\left\|\widetilde{\mathbf{S}}(y,\mathbf{M})^{-\frac{1}{2}}(\widetilde{\mathbf{S}}(y,\mathbf{M}) - \widetilde{\mathbf{S}}(x,\mathbf{M}))\widetilde{\mathbf{S}}(x,\mathbf{M})^{-\frac{1}{2}}\right\|_{\mathrm{op}} \leq 4r.$$

Now denote $\boldsymbol{\Delta}_C = \widetilde{\mathbf{C}}(y,\mathbf{M}) - \widetilde{\mathbf{C}}(x,\mathbf{M})$. We observe that the following inequality holds

$$\begin{aligned}\left\langle \boldsymbol{\Delta}_C\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}, \widetilde{\mathbf{C}}(y,\mathbf{M})^{-1}\boldsymbol{\Delta}_C\right\rangle &= \mathrm{tr}\big(\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}\boldsymbol{\Delta}_C\widetilde{\mathbf{C}}(y,\mathbf{M})^{-1}\boldsymbol{\Delta}_C\big) \\ &= \left\|\widetilde{\mathbf{C}}(y,\mathbf{M})^{-\frac{1}{2}}\boldsymbol{\Delta}_C\widetilde{\mathbf{C}}(x,\mathbf{M})^{-\frac{1}{2}}\right\|_{\mathrm{F}}^2 \geq 0.\end{aligned}$$

due to the property of trace. We compute

$$\begin{aligned}\left\|\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}\boldsymbol{\Delta}_C^2\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}\right\|_{\mathrm{op}} &\leq \left\|\boldsymbol{\Delta}_C\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}\right\|_{\mathrm{F}}^2 \\ &\leq \left\|\widetilde{\mathbf{C}}(y,\mathbf{M})^{-1}\boldsymbol{\Delta}_C + \boldsymbol{\Delta}_C\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}\right\|_{\mathrm{F}}^2 \\ &= \left\|\widetilde{\mathbf{C}}(y,\mathbf{M})^{-1}(\widetilde{\mathbf{C}}(y,\mathbf{M})^2 - \widetilde{\mathbf{C}}(x,\mathbf{M})^2)\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}\right\|_{\mathrm{F}}^2 \leq 16dr^2\end{aligned}$$

where the last line is due to Lemma 22. As a consequence, it holds that

$$\left\|\mathrm{sym}(\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}(\widetilde{\mathbf{C}}(y,\mathbf{M}) - \widetilde{\mathbf{C}}(x,\mathbf{M})))\right\|_{\mathrm{op}} \leq 4r + 16dr^2.$$

To prove the final property, we verify the Lipschitz condition required by Lemma 10. Consider any $u, v \in x + (r_{\max} + \sigma)\mathbf{M}^{-\frac{1}{2}} B_d \subseteq K$. We define $z_t = v + t(u - v)$ and obtain the following from Point 2 of Lemma 3:

$$(1 - r_{\max} - \sigma)^2 \mathbf{M} \preceq \nabla^2 \phi(z_t) \preceq (1 - r_{\max} - \sigma)^{-2} \mathbf{M}.$$

We compute the derivative of $\mathbf{S}$ along z_t as

$$\mathrm{D}\mathbf{S}(z_t)[h] = -\nabla^2 \phi(z_t)^{-1} \nabla^3 \phi(z_t)[h, \cdot, \cdot] \nabla^2 \phi(z_t)^{-1}.$$

The above Hessian comparison and 1-self-concordance imply

$$\begin{aligned} \left\| \mathrm{D}\mathbf{S}(z_t)[h] \right\|_{\mathbf{S}(x)^{-1}} &\leq \left\| \mathbf{M}^{\frac{1}{2}} \nabla^2 \phi(z_t)^{-\frac{1}{2}} \right\|_{\mathrm{op}}^2 \left\| \nabla^2 \phi(z_t)^{-\frac{1}{2}} \nabla^3 \phi(z_t)[h, \cdot, \cdot] \nabla^2 \phi(z_t)^{-\frac{1}{2}} \right\|_{\mathrm{op}} \\ &\leq \frac{2}{(1 - r_{\max} - \sigma)^2} \|h\|_{\nabla^2 \phi(z_t)} \leq \frac{2}{(1 - r_{\max} - \sigma)^3} \|h\|_{\mathbf{M}} \leq 20 \|h\|_{\mathbf{M}}, \end{aligned}$$

where the last inequality follows from $2(1 - r_{\max} - \sigma)^{-3} \leq 2 + 30(r_{\max} + \sigma) \leq 20$. Integrating over $t \in [0, 1]$ gives

$$\|\mathbf{S}(u) - \mathbf{S}(v)\|_{\mathbf{S}(x)^{-1}} \leq 20 \|u - v\|_{\mathbf{M}}.$$

Thus, $\mathbf{S}$ satisfies the Lipschitz condition in Lemma 10 with $G = 20$. By Point 2 of Lemma 10, we obtain

$$\left\| \widetilde{\mathbf{S}}(y, \mathbf{M}) - \widetilde{\mathbf{S}}(x, \mathbf{M}) - \mathrm{D}\widetilde{\mathbf{S}}(x, \mathbf{M})[y - x] \right\|_{\mathbf{S}(x)^{-1}} \leq \frac{10\sqrt{d}}{\sigma} \|y - x\|_{\mathbf{M}}^2. \tag{11}$$

We define $\boldsymbol{\Delta}_C := \widetilde{\mathbf{C}}(y, \mathbf{M}) - \widetilde{\mathbf{C}}(x, \mathbf{M}) - \mathrm{D}\widetilde{\mathbf{C}}(x, \mathbf{M})[y - x]$. Using $\widetilde{\mathbf{S}} = \widetilde{\mathbf{C}}^2$, we have

$$\begin{aligned} &\widetilde{\mathbf{S}}(y, \mathbf{M}) - \widetilde{\mathbf{S}}(x, \mathbf{M}) - \mathrm{D}\widetilde{\mathbf{S}}(x, \mathbf{M})[y - x] \\ &\quad = \widetilde{\mathbf{C}}(x, \mathbf{M}) \boldsymbol{\Delta}_C + \boldsymbol{\Delta}_C \widetilde{\mathbf{C}}(x, \mathbf{M}) + \left(\widetilde{\mathbf{C}}(y, \mathbf{M}) - \widetilde{\mathbf{C}}(x, \mathbf{M}) \right)^2 \\ &\quad = 2\mathrm{sym}(\widetilde{\mathbf{C}}(x, \mathbf{M}) \boldsymbol{\Delta}_C) + \left(\widetilde{\mathbf{C}}(y, \mathbf{M}) - \widetilde{\mathbf{C}}(x, \mathbf{M}) \right)^2. \end{aligned}$$

Rearranging yields

$$\begin{aligned} \mathrm{sym}\left(\boldsymbol{\Delta}_C \widetilde{\mathbf{C}}(x, \mathbf{M})^{-1} \right) &= \frac{1}{2} \widetilde{\mathbf{C}}(x, \mathbf{M})^{-1} \left(\widetilde{\mathbf{S}}(y, \mathbf{M}) - \widetilde{\mathbf{S}}(x, \mathbf{M}) - \mathrm{D}\widetilde{\mathbf{S}}(x, \mathbf{M})[y - x] \right) \widetilde{\mathbf{C}}(x, \mathbf{M})^{-1} \\ &\quad - \frac{1}{2} \left(\left(\widetilde{\mathbf{C}}(y, \mathbf{M}) - \widetilde{\mathbf{C}}(x, \mathbf{M}) \right) \widetilde{\mathbf{C}}(x, \mathbf{M})^{-1} \right)^\top \left(\left(\widetilde{\mathbf{C}}(y, \mathbf{M}) - \widetilde{\mathbf{C}}(x, \mathbf{M}) \right) \widetilde{\mathbf{C}}(x, \mathbf{M})^{-1} \right). \end{aligned}$$

While the first term is bounded in (11), the second term is bounded by $8d$, as an application of Lemma 22 and the fourth property. We conclude as

$$\left\| \mathrm{sym}\left(\boldsymbol{\Delta}_C \widetilde{\mathbf{C}}(x, \mathbf{M})^{-1} \right) \right\|_{\mathrm{op}} \leq \left(8d + \frac{20\sqrt{d}}{\sigma} \right) \|y - x\|_{\mathbf{M}}^2$$

by combining all the above computations. □

Lemma 24. *Let $\phi : K \subseteq \mathbb{R}^d \to \mathbb{R}$ be a 1-self-concordant barrier function. Let x, x' be reference points, $\mathbf{M} = \nabla^2\phi(x)$, $\mathbf{M}' = \nabla^2\phi(x')$ be the corresponding reference matrices, and $\sigma > 0$ be the smoothing scale. If $r := \|x - x'\|_{\mathbf{M}} \le r_{\max}$, $r_{\max} \le 1/5$, $\sigma < 1/5$, then the following property holds for $\widetilde{\mathbf{C}}(x, \mathbf{M})$ and $\widetilde{\mathbf{C}}(x, \mathbf{M}')$ defined as in Lemma 23,*

1. $\widetilde{\mathbf{S}}(x, \mathbf{M}) \approx_{4\sigma r} \widetilde{\mathbf{S}}(x, \mathbf{M}')$;
2. $\left\|\mathrm{sym}(\widetilde{\mathbf{C}}(x, \mathbf{M})^{-1}(\widetilde{\mathbf{C}}(x, \mathbf{M}) - \widetilde{\mathbf{C}}(x, \mathbf{M}')))\right\|_{\mathrm{op}} \le 2\sigma r + 10d\sigma^2 r^2$;
3. $\left\|\mathrm{sym}(\widetilde{\mathbf{C}}(x, \mathbf{M}')^{-1}(\widetilde{\mathbf{C}}(x, \mathbf{M}) - \widetilde{\mathbf{C}}(x, \mathbf{M}')))\right\|_{\mathrm{op}} \le 2\sigma r + 10d\sigma^2 r^2$.

Proof. It is easy to assert $\|x - x'\|_{\mathbf{M}'} \le r/(1-r) \le 1/4$, hence $\widetilde{\mathbf{C}}(x, \mathbf{M})$ and $\widetilde{\mathbf{C}}(x, \mathbf{M}')$ are immediately well-defined from Lemma 23.

Let $(\mathbf{M}^\dagger)^{-\frac{1}{2}} := \mathbf{M}^{-\frac{1}{2}}(\mathbf{M}^{-\frac{1}{2}}\mathbf{M}'\mathbf{M}^{-\frac{1}{2}})^{-\frac{1}{2}}$. By rotational invariance, $(\mathbf{M}^\dagger)^{-\frac{1}{2}}u$ and $\mathbf{M}'^{-\frac{1}{2}}u$ obey the same distribution as $(\mathbf{M}^\dagger)^{-\frac{1}{2}}((\mathbf{M}^\dagger)^{-\frac{1}{2}})^\top = \mathbf{M}'^{-1}$. Now, for any $u \in B_d$, it holds that

$$\begin{aligned}\left\|\sigma(\mathbf{M}^\dagger)^{-\frac{1}{2}}u - \sigma\mathbf{M}^{-\frac{1}{2}}u\right\|_{\nabla^2\phi(x+\sigma\mathbf{M}^{-\frac{1}{2}}u)} &\le \frac{\sigma}{1-\sigma}\left\|(\mathbf{M}^{-\frac{1}{2}}\mathbf{M}'\mathbf{M}^{-\frac{1}{2}})^{-\frac{1}{2}} - I\right\|_{\mathrm{op}} \\ &\le \frac{\sigma r}{(1-\sigma)(1-r)} \le \frac{2\sigma r}{1+2\sigma r}.\end{aligned}$$

where in the first inequality we compare the hessian between x and $x + \sigma\mathbf{M}^{-\frac{1}{2}}u$, and in the second we compare between x and x'. Also, the last inequality holds since $r + \sigma < 1/2$. We apply Lemma 3 again and obtain

$$(1+2\sigma r)^{-2}\mathbf{S}(x + \sigma\mathbf{M}^{-\frac{1}{2}}u) \preceq \mathbf{S}(x + \sigma(\mathbf{M}^\dagger)^{-\frac{1}{2}}u) \preceq (1+2\sigma r)^2\mathbf{S}(x + \sigma\mathbf{M}^{-\frac{1}{2}}u).$$

Taking expectations and using rotational invariance gives

$$(1+2\sigma r)^{-2}\widetilde{\mathbf{S}}(x, \mathbf{M}) \preceq \widetilde{\mathbf{S}}(x, \mathbf{M}') \preceq (1+2\sigma r)^2\widetilde{\mathbf{S}}(x, \mathbf{M}).$$

This is the first claim. It also implies following bounds:

$$\begin{aligned}\left\|\widetilde{\mathbf{C}}(x, \mathbf{M})^{-1}(\widetilde{\mathbf{S}}(x, \mathbf{M}') - \widetilde{\mathbf{S}}(x, \mathbf{M}))\widetilde{\mathbf{C}}(x, \mathbf{M})^{-1}\right\|_{\mathrm{op}} &\le 4\sigma r + 4\sigma^2 r^2, \\ \left\|\widetilde{\mathbf{C}}(x, \mathbf{M}')^{-1}(\widetilde{\mathbf{S}}(x, \mathbf{M}') - \widetilde{\mathbf{S}}(x, \mathbf{M}))\widetilde{\mathbf{C}}(x, \mathbf{M})^{-1}\right\|_{\mathrm{op}} &\le 4\sigma r,\end{aligned}$$

where the second inequality follows from the singular values of $\widetilde{\mathbf{C}}(x, \mathbf{M}')\widetilde{\mathbf{C}}(x, \mathbf{M})^{-1}$. We introduce the local notation $\boldsymbol{\Delta}_C = \widetilde{\mathbf{C}}(x, \mathbf{M}') - \widetilde{\mathbf{C}}(x, \mathbf{M})$. Since $\boldsymbol{\Delta}_C$ is symmetric, we have

$$\left\langle \boldsymbol{\Delta}_C\widetilde{\mathbf{C}}(x, \mathbf{M})^{-1}, \widetilde{\mathbf{C}}(x, \mathbf{M}')^{-1}\boldsymbol{\Delta}_C \right\rangle = \left\|\widetilde{\mathbf{C}}(x, \mathbf{M}')^{-\frac{1}{2}}\boldsymbol{\Delta}_C\widetilde{\mathbf{C}}(x, \mathbf{M})^{-\frac{1}{2}}\right\|_{\mathrm{F}}^2 \ge 0.$$

Consequently,

$$\begin{aligned}\left\|\widetilde{\mathbf{C}}(x, \mathbf{M})^{-1}\boldsymbol{\Delta}_C^2\widetilde{\mathbf{C}}(x, \mathbf{M})^{-1}\right\|_{\mathrm{op}} &\le \left\|\boldsymbol{\Delta}_C\widetilde{\mathbf{C}}(x, \mathbf{M})^{-1}\right\|_{\mathrm{F}}^2 \\ &\le \left\|\widetilde{\mathbf{C}}(x, \mathbf{M}')^{-1}\boldsymbol{\Delta}_C + \boldsymbol{\Delta}_C\widetilde{\mathbf{C}}(x, \mathbf{M})^{-1}\right\|_{\mathrm{F}}^2 \\ &= \left\|\widetilde{\mathbf{C}}(x, \mathbf{M}')^{-1}(\widetilde{\mathbf{S}}(x, \mathbf{M}') - \widetilde{\mathbf{S}}(x, \mathbf{M}))\widetilde{\mathbf{C}}(x, \mathbf{M})^{-1}\right\|_{\mathrm{F}}^2 \\ &\le 16d\sigma^2 r^2.\end{aligned}$$

Finally, using $\widetilde{\mathbf{S}} = \widetilde{\mathbf{C}}^2$, we obtain

$$\begin{aligned}2\mathrm{sym}(\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}\boldsymbol{\Delta}_C) = \widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}(\widetilde{\mathbf{S}}(x,\mathbf{M}') - \widetilde{\mathbf{S}}(x,\mathbf{M}))\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1} \\ - \widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}\boldsymbol{\Delta}_C^2\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}.\end{aligned}$$

As a result,

$$\left\|\mathrm{sym}(\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}\boldsymbol{\Delta}_C)\right\|_{\mathrm{op}} \le 2\sigma r + (8d+2)\sigma^2 r^2 \le 2\sigma r + 10d\sigma^2 r^2.$$

The same bound holds with $\widetilde{\mathbf{C}}(x,\mathbf{M}')^{-1}$ in place of $\widetilde{\mathbf{C}}(x,\mathbf{M})^{-1}$, by exchanging the reference matrices in the normalized bounds and applying the same Frobenius argument. □

By combining the above lemmas, when x and x' close, their corresponding Lewis weights are also close.

Corollary 25. *For any $\epsilon, \sigma \in (0, 1/10]$, two points x and x', if $\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}(x_i - x_i')\right\|_2 \le \epsilon$, $\tau(x) \approx_{40\epsilon} \tau(x')$ and $\widetilde{\mathbf{S}}(x, \nabla^2\Phi(x)) \approx_{5\epsilon} \widetilde{\mathbf{S}}(x', \nabla^2\Phi(x'))$, where $\tau(\cdot)$ denotes the preconditioned Lewis weight.*

Proof. As $x \approx_\epsilon x'$, we directly invoke Lemma 23 and Lemma 24 to get

$$\widetilde{\mathbf{S}}(x, \nabla^2\Phi(x)) \approx_{4\epsilon} \widetilde{\mathbf{S}}(x', \nabla^2\Phi(x)) \approx_{4\epsilon\sigma} \widetilde{\mathbf{S}}(x', \nabla^2\Phi(x')).$$

This is $\widetilde{\mathbf{S}}(x, \nabla^2\Phi(x)) \approx_{5\epsilon} \widetilde{\mathbf{S}}(x', \nabla^2\Phi(x'))$ with the choice of σ. By an application of Lemma 21, we obtain

$$\tau(x) = w\left(\widetilde{\mathbf{C}}(x, \nabla^2\Phi(x))\right) \approx_{40\epsilon} w\left(\widetilde{\mathbf{C}}(x', \nabla^2\Phi(x'))\right) = \tau(x')$$

and finish the proof. □

4.2 Stability of Variables

The potential decrease analysis requires us to bound the changes in the primal-dual variables, Lewis weights, and other quantities after a single call to ShortStep. We write

$$\tau = \tau(x), \qquad \mathbf{C} = \widetilde{\mathbf{C}}(x, \nabla^2\Phi(x)), \qquad \mu = \mu(x, s, \rho), \qquad \text{and,} \qquad \Gamma = \Gamma(\mu, x).$$

Moreover, let $\tau^{(\mathrm{new})}$, $\mathbf{C}^{(\mathrm{new})}$, $\mu^{(\mathrm{new})}$, and $\Gamma^{(\mathrm{new})}$ denote the corresponding quantities after the call to ShortStep. This subsection establishes bounds on their changes. To highlight the novel analysis of $\mathbf{C}$, we present its proof here and defer the remaining proofs to Appendix B. We will suppose (x, s, ρ) is ϵ-central for the rest of the section.

Stability of x and s. We begin with the primal and dual variables. Also, the norm bounds for deterministic direction vectors g, δ_1, and δ_2 in Algorithm 1 are also provided as a byproduct of primal analysis. Their proofs are postponed to Appendix B.1.

Lemma 26. *Let $(\overline{x}, \overline{\tau})$ be an α-approximation of (x, τ). Then the following properties hold for δ_x in Algorithm 1:*

- $\left\|\nabla^2\Phi(x)^{\frac{1}{2}}\mathbb{E}[\delta_x]\right\|_{2,\tau+\infty} \le (1 + O(C_{norm}^{-1} + \alpha + \beta))\beta$;
- $\left\|\nabla^2\Phi(x)^{\frac{1}{2}}\delta_x\right\|_{2,\infty} \lesssim \beta$ *w.h.p.*;

- $\left\|(\mathbb{E}\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\|_2)_{i\in[m]}\right\|_{\tau+\infty} \lesssim \beta$;
- $\left\|(\mathbb{E}\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\|_2^2)_{i\in[m]}\right\|_{\tau+\infty} \lesssim \beta^2$;

More over, it holds that $\|g\|_{2,\tau+\infty}, \|\delta_1\|_{2,\tau+\infty} \lesssim \beta$, $\|\delta_2\|_{2,\tau+\infty} \lesssim \alpha\beta$.

Corollary 27. *Let $(\overline{x},\overline{\tau})$ be an α-approximation of (x,τ). Then the following property holds for δ_s in Algorithm 1:*

$$\rho^{-1}\left\|\mathbf{T}^{-1}\nabla\Phi(x)^{-\frac{1}{2}}\delta_s\right\|_{\tau+\infty} \lesssim \beta$$

Stability of C. We next analyze the stability of $\mathbf{C}$. Due to the novelty of the analysis, we will present the lemma statement and its proof here.

Lemma 28. *Let $(\overline{x},\overline{\tau})$ be an α-approximation of (x,τ). Define $\mathbf{C} := \widetilde{\mathbf{C}}(x,\nabla^2\Phi(x))$, $\mathbf{C}^{(\mathrm{new})} := \widetilde{\mathbf{C}}(x^{(\mathrm{new})},\nabla^2\Phi(x^{(\mathrm{new})}))$ to be the smoothed function per Definition 11, and $\Delta_C = \mathbf{C}^{(\mathrm{new})} - \mathbf{C}$. Let*

$$l_i := \left\|\mathrm{sym}(\mathbf{C}_i^{-1}\boldsymbol{\Delta}_{C,i})\right\|_{\mathrm{op}}.$$

Then the following properties hold

- $\left\|\left(\left\|\mathbb{E}[\mathrm{sym}(\mathbf{C}_i^{-1}\boldsymbol{\Delta}_{C,i})]\right\|_{\mathrm{op}}\right)_{i\in[m]}\right\|_{\tau+\infty} \le (1+O(\alpha+\beta+C_{norm}^{-1}))\beta$;
- $\|l\|_\infty \lesssim \beta$ *w.h.p.*;
- $\left\|\mathbb{E}[l^2]\right\|_{\tau+\infty} \lesssim \beta^2$.

Proof. For the first claim, we expand the update $\boldsymbol{\Delta}_C$ as

$$\boldsymbol{\Delta}_C = \widetilde{\mathbf{C}}^{(\mathrm{new})} - \widetilde{\mathbf{C}} = \underbrace{\widetilde{\mathbf{C}}(x^{(\mathrm{new})},\mathbf{M}^{(\mathrm{new})}) - \widetilde{\mathbf{C}}(x^{(\mathrm{new})},\mathbf{M})}_{:=\boldsymbol{\Delta}_{\mathrm{local}}} + \underbrace{\widetilde{\mathbf{C}}(x^{(\mathrm{new})},\mathbf{M}) - \widetilde{\mathbf{C}}(x,\mathbf{M})}_{:=\boldsymbol{\Delta}_{\mathrm{drift}}}. \tag{12}$$

We will bound the local and drift norm separately. For the drift term, the relative change in i-th block satisfies

$$\begin{aligned}\left\|\mathbb{E}[\mathrm{sym}(\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{drift},i})]\right\|_{\mathrm{op}} \le{}& \left\|\mathrm{sym}(\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\mathrm{D}\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)[\mathbb{E}[\delta_{x,i}]])\right\|_{\mathrm{op}} \\ &+ \mathbb{E}\left[\left\|\mathrm{sym}\big(\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\big(\boldsymbol{\Delta}_{\mathrm{drift},i} - \mathrm{D}\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)[\delta_{x,i}]\big)\big)\right\|_{\mathrm{op}}\right].\end{aligned} \tag{13}$$

due to the triangle inequality and the linearity of expectation. Moreover, Lemma 26 indicates $\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}(x_i^{(\mathrm{new})} - x_i)\right\|_2 \lesssim \beta$. This allows us to invoke Point 3 in Lemma 23 to bound the first term in (13)

$$\left\|\mathrm{sym}(\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\mathrm{D}\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)[\mathbb{E}[\delta_{x,i}]])\right\|_{\mathrm{op}} \le (1+O(\beta+\sigma))\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\mathbb{E}[\delta_{x,i}]\right\|_2.$$

Similarly, Point 5 in the same lemma suggests the second term in (13) is bounded as

$$\left\|\mathrm{sym}\big(\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\big(\boldsymbol{\Delta}_{\mathrm{drift},i} - \mathrm{D}\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)[\delta_{x,i}]\big)\big)\right\|_{\mathrm{op}} \le O\left(\left(\frac{\sqrt{n_i}}{\sigma} + n_i\right)\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\right\|_2^2\right)$$

As a consequence, it holds that

$$\begin{aligned}\left\|\mathbb{E}[\mathrm{sym}(\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{drift},i})]\right\|_{\mathrm{op}} \leq (1+O(\beta+\sigma))\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\mathbb{E}[\delta_{x,i}]\right\|_2 \\ + O(n_i+\sqrt{n_i}/\sigma)\mathbb{E}\left[\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\right\|_2^2\right]\end{aligned}$$

For the local term in (12), the following identity holds

$$\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{local},i} = \widetilde{\mathbf{C}}_i(x_i^{(\mathrm{new})},\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{local},i} - \widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{drift},i}\widetilde{\mathbf{C}}_i(x_i^{(\mathrm{new})},\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{local},i}.$$

As a consequence, we compute

$$\begin{aligned}\left\|\mathrm{sym}(\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{local},i})\right\|_{\mathrm{op}} \leq \left\|\mathrm{sym}(\widetilde{\mathbf{C}}_i(x_i^{(\mathrm{new})},\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{local},i})\right\|_{\mathrm{op}} \\ + \left\|\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{drift},i}\right\|_{\mathrm{op}}\left\|\widetilde{\mathbf{C}}_i(x_i^{(\mathrm{new})},\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{local},i}\right\|_{\mathrm{op}}\end{aligned}$$

using Cauchy-Schwarz. In particular, it holds that

$$\left\|\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{drift},i}\right\|_{\mathrm{op}} \leq \sqrt{2n_i}\left\|\mathrm{sym}(\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{drift},i})\right\|_{\mathrm{op}} \lesssim \beta\sqrt{n_i}$$

due to Lemma 22, Lemma 26, and Point 4 in Lemma 23. Then the following holds

$$\begin{aligned}\left\|\mathrm{sym}(\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{local},i})\right\|_{\mathrm{op}} &\leq (1+O(n_i\beta))\left\|\mathrm{sym}\left(\widetilde{\mathbf{C}}_i(x_i^{(\mathrm{new})},\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{local},i}\right)\right\|_{\mathrm{op}} \\ &\lesssim \sigma\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\right\|_2\end{aligned} \tag{14}$$

where we use $n_i\beta \lesssim 1$, Lemma 24, and Lemma 22 again. We insert the above computation back to (12) and obtain

$$\begin{aligned}\left\|\left(\|\mathbb{E}[\mathrm{sym}(\mathbf{C}_i^{-1}\boldsymbol{\Delta}_{C,i})]\|_{\mathrm{op}}\right)_{i\in[m]}\right\|_{\tau+\infty} \leq &(1+O(\beta+\sigma))\left\|\nabla^2\Phi(x)^{\frac{1}{2}}\mathbb{E}[\delta_x]\right\|_{\tau+\infty} \\ &+O\left(\sigma\left\|\left(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\|_2]\right)_{i\in[m]}\right\|_{\tau+\infty}\right) \\ &+O\left((n_i+\sqrt{n_i}/\sigma)\left\|\left(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\|_2^2]\right)_{i\in[m]}\right\|_{\tau+\infty}\right) \\ \leq &(1+O(\alpha+\beta+C_{\mathrm{norm}}^{-1}))\beta\end{aligned}$$

due to Lemma 26, $n_i\beta = O(1)$, $\sigma = \theta/C = C_{\mathrm{norm}}^{-1}$, and also $\beta^2/\sigma = \beta/(C_\lambda\sqrt{\lambda}) \leq \beta/C_{\mathrm{norm}}$.

Now we establish the second claim. By Point 4 of Lemma 23, we have

$$\left\|\mathrm{sym}(\widetilde{\mathbf{C}}_i(x_i,\mathbf{M}_i)^{-1}\boldsymbol{\Delta}_{\mathrm{drift},i})\right\|_{\mathrm{op}} \lesssim (1+\beta n_i)\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\right\|_2.$$

Combined with the blockwise bound for local term in (14), we estimate

$$\|l\|_\infty \lesssim \max_{i\in[m]}(1+\beta n_i+2\sigma)\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\right\|_2 \lesssim \beta$$

by using the triangle inequality, Lemma 26, and our parameter choice. In a similar fashion, we also obtain

$$\left\|\mathbb{E}[l^2]\right\|_{\tau+\infty} \leq \left\|\left(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\|_2^2]\right)_{i\in[m]}\right\|_{\tau+\infty} \lesssim \beta^2.$$

□

We provide an alternative lemma to bound the stability of l under a weighted norm. Its proof is postponed to Appendix B.1.

Lemma 29. *Suppose the setting in Lemma 28. Define* $\mathbf{\Delta}_C = \mathbf{C}^{(\text{new})} - \mathbf{C}$, $\mathbf{C}_t = \mathbf{C} + t\mathbf{\Delta}_C$, $\tau_t = w(\mathbf{C}_t)$, *and* $\mathbf{P}_t = \mathbf{P}(\mathbf{T}_t^{1/2-1/p}\mathbf{C}_t\mathbf{A})$. *Then the following property holds for* $\mathbf{\Delta}_C$, $\mathbf{P}_t$, *and* l

$$\mathbb{E}\big[\|l\|_{\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t}\big] \lesssim \beta / C_{norm}.$$

Stability of τ and μ. On the basis of the previous bounds for x, s, and μ, we control the relative change of the Lewis weights and the residual error μ. In particular, unlike the scalar setting, the block Lewis-weight analysis must control the noncommutative matrix perturbations without introducing additional condition number. As a consequence, the scalar arguments do not directly extend to our setting. Although these extensions require substantial technical work, we defer the proofs for τ and μ to Appendix B.2 and Appendix B.3, respectively for better readability.

Lemma 30. *Let* $(\overline{x}, \overline{\tau})$ *be* α*-approximation of* (x, τ). *Define* $\tau = w(\mathbf{C})$, $\tau^{(\text{new})} = w(\mathbf{C}^{(\text{new})})$. *Then the following properties hold for* $\delta_\tau := \tau^{(\text{new})} - \tau$*:*

- $\big\|\mathbf{T}^{-1}\delta_\tau\big\|_\infty \lesssim \beta$ *w.h.p.;*
- $\big\|\mathbb{E}[(\mathbf{T}^{-1}\delta_\tau)^2]\big\|_{\tau+\infty} \lesssim \beta^2$*;*
- $\big\|\mathbf{T}^{-1}\mathbb{E}[\delta_\tau]\big\|_{\tau+\infty} \leq \sqrt{p/(2-p)} \cdot \beta + O((C_{norm}^{-1} + \alpha + \beta)\beta)$.

Lemma 31. *Let* $(\overline{x}, \overline{\tau})$ *be* α*-approximation of* (x, τ). *Define* μ, $\mu^{(\text{new})}$ *as*

$$\mu_i = \frac{s_i}{\rho\tau_i} + \nabla\phi_i(x_i), \qquad \mu_i^{(\text{new})} = \frac{s_i^{(\text{new})}}{\rho^{(\text{new})}\tau_i^{(\text{new})}} + \nabla\phi_i(x_i^{(\text{new})})$$

for any $i \in [m]$. *Then the following properties hold for* $\delta_\mu = \mu^{(\text{new})} - \mu$, g *and other quantities in Algorithm 1*

- $\big\|\nabla^2\Phi(x)^{-1/2}\mathbb{E}[\delta_\mu] - g\big\|_{2,\tau+\infty} \leq \sqrt{p/(2-p)} \cdot \beta + O((C_{norm}^{-1} + \alpha + \beta)\beta)$*;*
- $\big\|\big(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{-1/2}\delta_{\mu,i}\|_2^2]\big)_{i\in[m]}\big\|_{\tau+\infty} \lesssim \beta^2$*;*
- $\big\|\nabla^2\Phi(x)^{-\frac{1}{2}}\delta_\mu\big\|_{2,\infty} \lesssim \beta$.

4.3 Bounds of Bregman Metric

We will put together the stability bounds in Section 4.2 and analyze the maximal change of Bregman centrality metric Γ.

Lemma 32. *Let* $(\overline{x}, \overline{\tau})$ *be* α*-approximation of* (x, τ). *Define* $\Gamma = \Gamma(\mu, x)$, $\Gamma^{(\text{new})} = \Gamma(\mu^{(\text{new})}, x^{(\text{new})})$. *Let* $\|\Gamma\|_\infty \leq \epsilon^2$, *then the following property holds for* $\delta_\Gamma := \Gamma^{(\text{new})} - \Gamma$*:*

$$\big\|\delta_\Gamma\big\|_\infty \lesssim \beta\epsilon + \beta^2.$$

Proof. For this proof and the proof of Lemma 34 only, define $x_{t,i} = x_i + t\delta_{x,i}$ and $\mu_{t,i} = \mu_i + t\delta_{\mu,i}$ for any $t \in [0,1]$ and $i \in [m]$. We write with compact notation for dual variables $a_{t,i} = \nabla\phi_i(x_{t,i}) - \mu_{t,i}$, $b_{t,i} = \nabla\phi_i(x_{t,i})$. Due to the frequent use of the primal-dual derivative identities like

$$\nabla\phi^*(\cdot) = (\nabla\phi)^{-1}(\cdot), \qquad \nabla^2\phi^*(\cdot) = [\nabla^2\phi(\nabla\phi^*(\cdot)))]^{-1},$$

for any ϕ, we do not refer to Lemma 5 each time they are used. Additionally, define

$$\Delta^*_{\phi_{t,i}} := \nabla\phi_i^*(a_{t,i}) - \nabla\phi_i^*(b_{t,i}) = (\nabla\phi_i)^{-1}(a_{t,i}) - x_{t,i}$$

and

$$\mathrm{SC}^*_{t,i} := \Delta^*_{\phi_{t,i}} + \nabla^2\phi_i^*(b_{t,i})\mu_{t,i} = \Delta^*_{\phi_{t,i}} + \nabla^2\phi_i(x_{t,i})^{-1}\mu_{t,i}.$$

When $t = 0$, we simply use $\Delta^*_{\phi_i}$ and SC^*_i. Now with a fixed $i \in [m]$, we temporarily omit the index subscript and define $g(t) = 2D_{\phi^*}(a_t, b_t)$. With the standard second-order Taylor expansion, it holds that

$$\Gamma^{(\mathrm{new})} - \Gamma = g(1) - g(0) = g'(0) + \frac{1}{2}g''(\xi). \tag{15}$$

for some $\xi \in [0,1]$. The first-order time derivative of $g(t)$ is

$$g'(t) = 2\left\langle \nabla\phi^*(a_t) - \nabla\phi^*(b_t), \nabla^2\phi(x_t)\delta_x - \delta_\mu \right\rangle + 2\left\langle \nabla^2\phi^*(b_t)\mu_t, \nabla^2\phi(x_t)\delta_x \right\rangle$$

due to the definition of a_t, b_t and the Bregman divergence. When $t = 0$, it simplifies to

$$g'(0) = 2\langle \mathrm{SC}^*, \nabla^2\phi(x)\delta_x\rangle - 2\langle \Delta^*_\phi, \delta_\mu\rangle$$

by simple rearranging. $g'(t)$ is written alternatively as

$$g'(t) = 2\langle \Delta^*_{\phi_t}, \nabla^2\phi(x_t)\delta_x - \delta_\mu\rangle + 2\langle \mu_t, \delta_x\rangle. \tag{16}$$

We compute the second-order derivative based on (16)

$$\begin{aligned}
\frac{1}{2}g''(t) =& \langle \nabla^2\phi^*(a_t)\nabla^2\phi(x_t)\delta_x - \delta_x, \nabla^2\phi(x_t)\delta_x - \delta_\mu\rangle + \langle \delta_\mu, \delta_x\rangle \\
&- \langle \nabla^2\phi^*(a_t)\delta_\mu, \nabla^2\phi(x_t)\delta_x - \delta_\mu\rangle + \nabla^3\phi(x_t)[\delta_x, \delta_x, \Delta^*_{\phi_t}] \\
=& \langle \nabla^2\phi^*(a_t)\delta_\mu, \delta_\mu\rangle - 2\langle \nabla^2\phi^*(a_t)\delta_\mu, \nabla^2\phi(x_t)\delta_x\rangle + 2\langle \delta_\mu, \delta_x\rangle \\
&+ \langle \nabla^2\phi^*(a_t)\nabla^2\phi(x_t)\delta_x - \delta_x, \nabla^2\phi(x_t)\delta_x\rangle + \nabla^3\phi(x_t)[\delta_x, \delta_x, \Delta^*_{\phi_t}]
\end{aligned}$$

where we use the following identity

$$\left(\Delta^*_{\phi_t}\right)' = \nabla^2\phi^*(a_t)\left(\nabla^2\phi(x_t)\delta_x - \delta_\mu\right) - \delta_x.$$

Let $\xi \in [0,1]$ be some variable given by the second-order Taylor expansion. As a consequence of the above computation, we obtain

$$\begin{aligned}
\Gamma^{(\mathrm{new})} - \Gamma = &\underbrace{2\langle \Delta^*_\phi, -\delta_\mu\rangle}_{L_1} + \underbrace{2\langle \mathrm{SC}^*, \nabla^2\phi(x)\delta_x\rangle}_{L_2} + \underbrace{\langle \nabla^2\phi^*(a_\xi)\delta_\mu, \delta_\mu\rangle}_{Q_1} + \underbrace{2\langle \nabla^2\phi^*(a_\xi)\delta_\mu, -\nabla^2\phi(x_\xi)\delta_x\rangle}_{Q_2} \\
&+ \underbrace{2\langle \delta_\mu, \delta_x\rangle}_{Q_3} + \underbrace{\langle \nabla^2\phi^*(a_\xi)\nabla^2\phi(x_\xi)\delta_x - \delta_x, \nabla^2\phi(x_\xi)\delta_x\rangle}_{Q_4} + \underbrace{\nabla^3\phi(x_\xi)[\delta_x, \delta_x, \Delta^*_{\phi_\xi}]}_{Q_5}.
\end{aligned} \tag{17}$$

In order to bound $|\Gamma_i^{(\text{new})} - \Gamma_i|$, we instantiate (17) for the i-th block and bound each term on the RHS. We retain the subscript i and start with $L_{1,i}$. We control the local norm of $\Delta^*_{\phi_i}$

$$\|\Delta^*_{\phi_i}\|_{\nabla^2\phi_i^*(b_i)^{-1}} \le 2\|\mu_i\|_{\nabla^2\phi_i^*(\nabla\phi(x))} = 2\|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}}. \tag{18}$$

due to the primal-dual identity and Point 1 in Lemma 6. It follows that

$$L_{1,i} = \langle \Delta^*_{\phi_i}, -\delta_{\mu,i}\rangle \le \|\delta_{\mu,i}\|_{\nabla^2\phi_i(x_i)^{-1}}\|\Delta^*_{\phi,i}\|_{\nabla^2\phi^*(b_i)^{-1}} \le 2\|\delta_{\mu,i}\|_{\nabla^2\phi_i(x_i)^{-1}}\|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}}$$

by Cauchy-Schwarz. For the term $L_{2,i}$, it requires the bound of SC_i

$$\|\mathrm{SC}_i^*\|_{\nabla^2\phi_i^*(b_i)^{-1}} \le 2\|\mu_i\|^2_{\nabla^2\phi_i^*(\nabla\phi(x))} = 2\|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}}. \tag{19}$$

as a consequence of Point 2 in Lemma 6. We estimate

$$L_{2,i} = \langle \mathrm{SC}_i^*, \nabla^2\phi_i(x_i)\delta_{x,i}\rangle \le \|\mathrm{SC}_i^*\|_{\nabla^2\phi_i^*(b_i)^{-1}}\|\delta_{x,i}\|_{\nabla^2\phi_i(x_i)} \le 2\|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}}\|\delta_{x,i}\|_{\nabla^2\phi_i(x_i)}.$$

We next turn to the quadratic terms in (17). Before this, we estimate the local norm of $\mu_{\xi,i}$ and $\nabla\phi_i^*(a_{\xi,i})$. Triangle inequality allows us to bound the first term

$$\|\mu_{\xi,i}\|_{\nabla^2\phi_i(x_i)^{-1}} \lesssim \|\delta_{\mu,i}\|_{\nabla^2\phi_i(x_i)^{-1}} + \|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}} \lesssim \epsilon + \beta, \tag{20}$$

and the last step is due to Lemma 31 and Lemma 8, as $\|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}} \le 2\Gamma_i \le 2\epsilon^2$. Consequently, we bound the second term

$$\|\nabla^2\phi_i^*(a_{\xi,i})\|_{\nabla^2\phi_i^*(b_{\xi,i})^{-1}} \le 1 + O\big(\|\Delta^*_{\phi_{\xi,i}}\|_{\nabla^2\phi_i(x_{\xi,i})}\big) \le 1 + O\big(\|\mu_{\xi,i}\|_{\nabla^2\phi_i(x_{\xi,i})^{-1}}\big) \le 2, \tag{21}$$

where we use arguments similar to (18) and (20).

For the term $Q_{1,i}$, it holds that

$$Q_{1,i} = \langle \nabla^2\phi_i^*(a_{\xi,i})\delta_{\mu,i}, \delta_{\mu,i}\rangle \lesssim \|\delta_{\mu,i}\|^2_{\nabla^2\phi_i(x_i)^{-1}}\|\nabla^2\phi_i^*(a_{\xi,i})\|_{\nabla^2\phi_i^*(b_{\xi,i})^{-1}} \lesssim \|\delta_{\mu,i}\|^2_{\nabla^2\phi_i(x_i)^{-1}}$$

where we use Point 2 in Lemma 3 and (20). Moreover, it requires the bound

$$\|x_{\xi,i} - x_i\|_{\nabla^2\phi_i(x_i)} = \xi\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\| \lesssim \beta$$

as a consequence of Lemma 26. For the term $Q_{2,i}$, it holds that

$$\begin{aligned} Q_{2,i} &= 2\langle \nabla^2\phi_i^*(a_{\xi,i})\delta_{\mu,i}, -\nabla^2\phi_i(x_{\xi,i})\delta_{x,i}\rangle \\ &\lesssim \|\nabla^2\phi_i^*(a_{\xi,i})\|_{\nabla^2\phi_i^*(b_{\xi,i})^{-1}}\left(\|\delta_{\mu,i}\|^2_{\nabla^2\phi_i(x_i)^{-1}} + \|\delta_{x,i}\|^2_{\nabla^2\phi_i(x_i)}\right) \\ &\lesssim \|\delta_{\mu,i}\|^2_{\nabla^2\phi_i(x_i)^{-1}} + \|\delta_{x,i}\|^2_{\nabla^2\phi_i(x_i)} \end{aligned}$$

due to the arguments identical to that of $Q_{1,i}$. For the term $Q_{3,i}$, it simply holds that

$$Q_{3,i} = 2\langle \delta_{\mu,i}, \delta_{x,i}\rangle \le 4\|\delta_{\mu,i}\|^2_{\nabla^2\phi_i(x_i)^{-1}} + 4\|\delta_{x,i}\|^2_{\nabla^2\phi_i(x_i)}$$

because of the Young's inequality. For the term $Q_{4,i}$, it holds that

$$\begin{aligned} Q_{4,i} &= \langle \nabla^2\phi_i^*(a_{\xi,i})\nabla^2\phi_i(x_{\xi,i})\delta_{x,i} - \delta_{x,i}, \nabla^2\phi_i(x_{\xi,i})\delta_{x,i}\rangle \\ &\lesssim \|\delta_{x,i}\|^2_{\nabla^2\phi_i(x_i)}\|\nabla^2\phi_i^*(a_{\xi,i}) - \nabla^2\phi_i^*(b_{\xi,i})\|_{\nabla^2\phi_i^*(b_{\xi,i})^{-1}} \end{aligned}$$

$$
\begin{aligned}
&= \|\delta_{x,i}\|^2_{\nabla^2\phi_i(x_i)} \|\nabla^2\phi_i(\nabla\phi_i^*(a_{\xi,i}))^{-1} - \nabla^2\phi_i(\nabla\phi_i^*(b_{\xi,i}))^{-1}\|_{\nabla^2\phi_i(x_{\xi,i})} \\
&\lesssim \|\delta_{x,i}\|^2_{\nabla^2\phi_i(x_i)} \|\Delta^*_{\phi_{\xi,i}}\|_{\nabla^2\phi_i^*(b_{\xi,i})^{-1}} \lesssim \|\delta_{x,i}\|^2_{\nabla^2\phi_i(x_i)} \|\mu_{\xi,i}\|_{\nabla^2\phi_i(x_i)^{-1}} \\
&\lesssim \|\delta_{x,i}\|^2_{\nabla^2\phi_i(x_i)} \left(\|\delta_{\mu,i}\|_{\nabla^2\phi_i(x_i)^{-1}} + \|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}}\right)
\end{aligned}
$$

also due to the arguments identical to that of $Q_{1,i}$. For the last term $Q_{5,i}$, we use 1-self-concordance and obtain

$$
\begin{aligned}
|Q_{5,i}| = \left|\nabla^3\phi_i(x_{\xi,i})\left[\delta_{x,i}, \delta_{x,i}, \Delta^*_{\phi_{\xi,i}}\right]\right| &\lesssim \|\delta_{x,i}\|^2_{\nabla^2\phi_i(x_{\xi,i})} \|\Delta^*_{\phi_{\xi,i}}\|_{\nabla^2\phi_i^*(b_{\xi,i})^{-1}} \\
&\lesssim \|\delta_{x,i}\|^2_{\nabla^2\phi_i(x_i)} \|\mu_{\xi,i}\|_{\nabla^2\phi_i(x_i)^{-1}} \\
&\lesssim \|\delta_{x,i}\|^2_{\nabla^2\phi_i(x_i)} \left(\|\delta_{\mu,i}\|_{\nabla^2\phi_i(x_i)^{-1}} + \|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}}\right).
\end{aligned}
$$

As a result, we obtain

$$
|\Gamma_i^{(\text{new})} - \Gamma_i| \le 2\sum_{j=1}^{2} |L_{j,i}| + \sum_{j=1}^{5} |Q_{j,i}| \le O(\beta\epsilon + \beta\epsilon^2 + \beta^2\epsilon + \beta^2) \le \beta\epsilon + \beta^2
$$

by considering all the parts above, Lemma 26, Lemma 31, and $\|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}} \lesssim \epsilon^2$. The final bound holds because the constant is absorbed by β and ϵ when C is chosen sufficiently large. □

The next lemma bound the maximal difference between the ideal Γ and the Bregman metric corresponding to the robust variables (x, τ, μ).

Lemma 33. *Let $(\overline{x}, \overline{\tau})$ be α-approximation of (x, τ), and $\overline{\mu}$ be 10α-approximation of μ. Define $\Gamma = \Gamma(\mu, x)$ and $\overline{\Gamma} = \Gamma(\overline{\mu}, \overline{x})$. If $\|\Gamma\|_\infty \le \epsilon^2$, then the following property holds for Γ and $\overline{\Gamma}$*

$$
\left|\Gamma_i - \overline{\Gamma}_i\right| \le 100\alpha\epsilon + 800\alpha^2.
$$

Proof. For this proof only, define compact dual variable for any $t \in [0, 1]$ and $i \in [m]$

$$
a_{t,i} = \nabla\phi_i(x_i) - t\mu_i, \qquad \overline{a}_{t,i} = \nabla\phi_i(\overline{x}_i) - t\overline{\mu}_i,
$$

and the corresponding primal variables $u_{t,i}$, $\overline{u}_{t,i}$ such that

$$
a_{t,i} = \nabla\phi_i(u_{t,i}), \qquad \overline{a}_{t,i} = \nabla\phi_i(\overline{u}_{t,i}).
$$

By the definition of the Bregman divergence, Γ_i has the following expansion

$$
\Gamma_i = 2\phi_i^*(a_{1,i}) - 2\phi_i^*(a_{0,i}) + 2\left\langle\nabla\phi_i^*(a_{0,i}), \mu_i\right\rangle = 2\int_0^1 (1-t)\left\langle\nabla^2\phi_i^*(a_{t,i}), \mu_i\mu_i^\top\right\rangle \mathrm{d}t.
$$

Similarly, we can write for $\overline{\Gamma}_i$

$$
\overline{\Gamma}_i = 2\int_0^1 (1-t)\left\langle\nabla^2\phi_i^*(\overline{a}_{t,i}), \overline{\mu}_i\overline{\mu}_i^\top\right\rangle \mathrm{d}t.
$$

Therefore the absolute difference between Γ_i and $\overline{\Gamma}_i$ is split as

$$
\left|\Gamma_i - \overline{\Gamma}_i\right| \le 2\underbrace{\left|\int_0^1 (1-t)\left\langle\left(\nabla^2\phi_i^*(a_{t,i}) - \nabla^2\phi_i^*(\overline{a}_{t,i})\right), \mu_i\mu_i^\top\right\rangle \mathrm{d}t\right|}_{\text{(I)}} + 2\underbrace{\left|\int_0^1 (1-t)\langle\nabla^2\phi_i^*(\overline{a}_{t,i}), \mu_i\mu_i^\top - \overline{\mu}_i\overline{\mu}_i^\top\rangle \mathrm{d}t\right|}_{\text{(II)}}.
$$

The two terms on the above RHS will be bounded separately. We claim the following inequality

$$\left\|\nabla^2\phi_i^*(a_{t,i})\right\|_{\nabla^2\phi_i^*(a_{0,i})^{-1}} \le 2, \qquad \left\|\nabla^2\phi_i^*(\overline{a}_{t,i})\right\|_{\nabla^2\phi_i^*(\overline{a}_{0,i})^{-1}} \le 2. \tag{22}$$

is true for any $t \in [0,1]$, $i \in [m]$, and postpone its proof to the end. We proceed to bound (I). By duality, Lemma 6, and (22), we compute

$$\begin{aligned}\|\nabla\phi_i^*(a_{t,i}) - \nabla\phi_i^*(\overline{a}_{t,i})\|_{\nabla^2\phi^*(a_{t,i})^{-1}} &\le \|a_{t,i} - \overline{a}_{t,i}\|_{\nabla^2\phi^*(a_{t,i})} \\ &\le \|\nabla^2\phi^*(a_{t,i})^{\frac{1}{2}}\|_{\nabla^2\phi^*(a_{0,i})^{-\frac{1}{2}}}\|a_{t,i} - \overline{a}_{t,i}\|_{\nabla^2\phi_i(x_i)^{-1}}.\end{aligned}$$

In particular, the last factor is bounded as

$$\begin{aligned}\|a_{t,i} - \overline{a}_{t,i}\|_{\nabla^2\phi_i(x_i)^{-1}} &\le \|\nabla\phi_i(x_i) - \nabla\phi_i(\overline{x}_i)\|_{\nabla^2\phi_i(\overline{x}_i)^{-1}} + \|\mu_i - \overline{\mu}_i\|_{\nabla^2\phi_i(\overline{x}_i)^{-1}} \\ &\le \left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}(x_i - \overline{x}_i)\right\|_2 + \|\mu_i - \overline{\mu}_i\|_{\nabla^2\phi_i(\overline{x}_i)^{-1}} \le 12\alpha.\end{aligned}$$

where we use Point 3 in Lemma 3, and the approximation ratio of $\overline{x}$ and $\overline{\mu}$. As a consequence, we repeatedly use the primal-dual derivative identities from Lemma 5 and obtain

$$\begin{aligned}\text{(I)} \le& \int (1-t)\left\|\nabla^2\phi_i^*(a_{t,i}) - \nabla^2\phi_i^*(\overline{a}_{t,i})\right\|_{\nabla^2\phi_i^*(a_{t,i})^{-1}}\|\nabla^2\phi^*(a_{t,i})\|_{\nabla^2\phi^*(a_{0,i})^{-1}}\|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}}\mathrm{d}t \\ \le& 3\int (1-t)\left\|\nabla^2\phi_i(u_{t,i})^{\frac{1}{2}}\nabla^2\phi_i(\overline{u}_{t,i})^{-1}\nabla^2\phi_i(u_{t,i})^{\frac{1}{2}} - \mathbf{I}\right\|_{\mathrm{op}}\|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}}\mathrm{d}t \\ \le& 6\int (1-t)\left\|\nabla\phi_i^*(a_{t,i}) - \nabla\phi_i^*(\overline{a}_{t,i})\right\|_{\nabla^2\phi^*(a_{t,i})^{-1}}\|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}}\mathrm{d}t \\ \le& 12\int (1-t)\|a_{t,i} - \overline{a}_{t,i}\|_{\nabla^2\phi^*(a_{t,i})}\|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}}\mathrm{d}t\end{aligned}$$

where the second line is due to (22), the third line is due to Point 2 in Lemma 3, and the last line comes from the above computation. Moreover, by Lemma 8, we have $\|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}} = \gamma_i \le 2\Gamma_i \le 2\epsilon^2$. As a consequence, it holds that (I) $\le 288\alpha\epsilon^2$. We proceed and split (II) as following

$$\text{(II)} \le \|\mu_i - \overline{\mu}_i\|^2_{\nabla^2\phi_i(\overline{x}_i)^{-1}} + 2\|\mu_i - \overline{\mu}_i\|_{\nabla^2\phi_i(\overline{x}_i)^{-1}}\|\mu_i\|_{\nabla^2\phi_i(\overline{x}_i)^{-1}}.$$

by using (22). By Hessian comparison and robust approximation, the first term in the above RHS is bounded as

$$\|\mu_i - \overline{\mu}_i\|^2_{\nabla^2\phi_i(\overline{x}_i)^{-1}} \le 2\|\mu_i - \overline{\mu}_i\|^2_{\nabla^2\phi_i(x_i)^{-1}} \le 200\alpha^2.$$

For the second term, we use Cauchy-Schwarz and compute

$$\|\mu_i - \overline{\mu}_i\|_{\nabla^2\phi_i(x_i)^{-1}}\|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}} \le 11\alpha\epsilon$$

Therefore (II) $\le 45\alpha\epsilon + 200\alpha^2$. As a result, we obtain

$$\left|\Gamma_i - \overline{\Gamma}_i\right| \le 100\alpha\epsilon + 800\alpha^2.$$

It remains to prove (22) for any i and t

$$\begin{aligned}\|\nabla^2\phi_i^*(a_{t,i})\|_{\nabla^2\phi_i^*(a_{0,i})^{-1}} &= \|\nabla^2\phi_i(u_{0,i})\|_{\nabla^2\phi_i(u_{t,i})^{-1}} \\ &\le 1 + 3\|\nabla\phi_i^*(a_{t,i}) - \nabla\phi_i^*(a_{0,i})\|_{\nabla^2\phi_i^*(a_{0,i})^{-1}} \\ &\le 1 + 6\|a_{t,i} - a_{0,i}\|_{\nabla^2\phi_i^*(a_{0,i})} = 1 + 6t\|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}} \le 1 + 12\epsilon \le 2\end{aligned}$$

due to Lemma 5, and upper bound for $\|\gamma\|_\infty$. The robust estimate follows identically for barred variables as $\|\nabla^2\phi^*(\overline{a}_{t,i})\|_{\nabla^2\phi^*(\overline{a}_{0,i})^{-1}} \le 2$. □

4.4 Potential Descent Analysis

In this subsection, we will prove the one-step descent property of the potential function, i.e. the target Lemma 20. The following lemma characterizes the change of potential function using the dual norm of potential gradients.

Lemma 34 (Potential change)**.** *Let $(\overline{x}, \overline{\tau})$ be α-approximation of (x, τ), $\overline{\mu}$ be 10α-approximation of μ, and $\|\Gamma\|_\infty \leq \epsilon^2$. Then the following property holds for g, γ, μ, $\mu^{(\mathrm{new})}$, x, and $x^{(\mathrm{new})}$ in Algorithm 1*

$$\begin{aligned}\mathbb{E}[\Psi(\mu^{(\mathrm{new})}, x^{(\mathrm{new})})] \leq \Psi(\mu, x) + \left\langle \nabla_\mu^{\mathrm{loc}} \Psi(\mu, x), g \right\rangle \\ + (1 - \theta/3)\beta \|\nabla_\mu^{\mathrm{loc}} \Psi(\mu, x)\|_{2,\tau+\infty}^* + O(\beta^2 + \lambda\beta^2\epsilon^2)\|\nabla_\Gamma \Psi(\mu, x)\|_{\tau+\infty}^*.\end{aligned}$$

Proof. In this proof, we reuse all the local notations in the proof of Lemma 32. We apply Lemma 4.10 in [LPT+25] with the update of each Γ_i and obtain the following inequality, using the notations and (17) from Lemma 32

$$\mathbb{E}[\Psi(\mu^{(\mathrm{new})}, x^{(\mathrm{new})})] \leq \Psi(\mu, x) + \sum_{i=1}^m \psi'(\Gamma_i) \left(2 \sum_{j=1}^2 \mathbb{E}[L_{j,i}] + \sum_{j=1}^5 \mathbb{E}[Q_{j,i}] \right) \tag{23}$$

$$+ 7\lambda e^{\lambda \|\delta_\Gamma\|_\infty} \sum_{i=1}^m \psi'(\Gamma_i) \left(4 \sum_{j=1}^2 \mathbb{E}[L_{j,i}^2] + \sum_{j=1}^5 \mathbb{E}[Q_{j,i}^2] \right). \tag{24}$$

due to inequality $(\sum_{i=1}^n a_i)^2 \leq n \sum_{i=1}^n a_i^2$ and $\psi''(\cdot) = \lambda \psi'(\cdot)$ for $\psi(\cdot) = \exp(\lambda(\cdot))$. It requires to analyze each term in (23) and (24). We first bound the terms in the L-series: for the $L_{1,i}$ term in (23), it holds that

$$\begin{aligned}2 \sum_{i=1}^m \psi'(\Gamma_i) \mathbb{E}[L_{1,i}] &= -\sum_{i=1}^m 2\psi'(\Gamma_i) \langle \nabla^2 \phi_i(x_i)^{\frac{1}{2}} \Delta_{\phi_i}^*, g_i \rangle + 2\psi'(\Gamma_i) \langle \Delta_{\phi_i}^*, \nabla^2 \phi_i(x_i)^{\frac{1}{2}} g_i - \mathbb{E}[\delta_{\mu,i}] \rangle \\ &= \sum_{i=1}^m \left\langle \left(\nabla_\mu^{\mathrm{loc}} \Psi(\mu, x) \right)_i, g_i \right\rangle + \left\langle \left(\nabla_\mu^{\mathrm{loc}} \Psi(\mu, x) \right)_i, g_i - \nabla^2 \phi_i(x_i)^{-\frac{1}{2}} \mathbb{E}[\delta_{\mu,i}] \right\rangle \\ &\leq \left\langle \nabla_\mu^{\mathrm{loc}} \Psi(\mu, x), g \right\rangle + \left\| \nabla^2 \Phi(x)^{\frac{1}{2}} \nabla_\mu \Psi(\mu, x) \right\|_{2,\tau+\infty}^* \left\| \nabla^2 \Phi(x)^{-\frac{1}{2}} \mathbb{E}[\delta_\mu] - g \right\|_{2,\tau+\infty} \\ &\leq \left\langle \nabla_\mu^{\mathrm{loc}} \Psi(\mu, x), g \right\rangle + \left(p\beta + O(\alpha + \beta + C_{\mathrm{norm}}^{-1})\beta \right) \left\| \nabla_\mu^{\mathrm{loc}} \Psi(\mu, x) \right\|_{2,\tau+\infty}^* \\ &\leq \left\langle \nabla_\mu^{\mathrm{loc}} \Psi(\mu, x), g \right\rangle + (1 - \theta/2)\beta \left\| \nabla_\mu^{\mathrm{loc}} \Psi(\mu, x) \right\|_{2,\tau+\infty}^*\end{aligned}$$

where the second line comes from the Cauchy-Schwarz and identity

$$\left(\nabla_\mu \Psi(\mu, x) \right)_i = \psi'(\Gamma_i) \nabla_{\mu_i} \Gamma_i = -2\psi'(\Gamma_i) \Delta_{\phi_i}^* \tag{25}$$

the third line is due to the the dual norm definition, the fourth line is due to Lemma 31, and the final line comes from the fact

$$\sqrt{\frac{p}{2-p}} + O(\alpha + \beta + C_{\mathrm{norm}}^{-1}) = 1 - \frac{2\theta}{3} + O(\theta/C) \leq 1 - \frac{\theta}{2}.$$

when our constant C is sufficiently large. For the $L_{2,i}$ term in (23), it holds that

$$\begin{aligned}
2\sum_{i=1}^m \psi'(\Gamma_i)\mathbb{E}[L_{2,i}] =& 2\sum_{i=1}^m \psi'(\Gamma_i)\langle \nabla^2\phi_i(x_i)^{\frac{1}{2}}\mathrm{SC}_i^*, \nabla^2\phi_i(x_i)^{\frac{1}{2}}\mathbb{E}[\delta_{x,i}]\rangle \\
\le& 4\sum_{i=1}^m \psi'(\Gamma_i)\big\|\Delta_{\phi_i}^*\big\|_{\nabla^2\phi_i(x_i)}\big\|\mu_i\big\|_{\nabla^2\phi_i(x_i)^{-1}}\big\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\mathbb{E}[\delta_{x,i}]\big\|_2 \\
=& 2\sum_{i=1}^m \big\|\big(\nabla_\mu^{\mathrm{loc}}\Psi(\mu,x)\big)_i\big\|_2\big\|\mu_i\big\|_{\nabla^2\phi_i(x_i)^{-1}}\big\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\mathbb{E}[\delta_{x,i}]\big\|_2 \\
\le& 4\epsilon\sum_{i=1}^m \big\|\big(\nabla_\mu^{\mathrm{loc}}\Psi(\mu,x)\big)_i\big\|_2\big\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\mathbb{E}[\delta_{x,i}]\big\|_2 \\
\le& 4\epsilon\big\|\nabla_\mu^{\mathrm{loc}}\Psi(\mu,x)\big\|_{2,\tau+\infty}^*\big\|\nabla^2\Phi(x)^{\frac{1}{2}}\mathbb{E}[\delta_x]\big\|_{2,\tau+\infty} \le \frac{8\theta}{C}\beta\big\|\nabla_\mu^{\mathrm{loc}}\Psi(\mu,x)\big\|_{2,\tau+\infty}^*
\end{aligned}$$

where the second line is due to Point 2 in Lemma 6, the middle equality is due to (25), the second last one is due to the fact $\|\gamma\|_\infty \le 2\|\Gamma\|_\infty \le 2\epsilon^2$, and the last line is due to the block duality norm definition, Lemma 26, and $\epsilon = \theta/C$. For the $L_{1,i}^2$ term in (24), we compute

$$\begin{aligned}
\sum_{i=1}^m \psi'(\Gamma_i)\mathbb{E}[L_{1,i}^2] &= \sum_{i=1}^m \psi'(\Gamma_i)\mathbb{E}\big[\langle \Delta_{\phi_i}^*, \delta_{\mu,i}\rangle^2\big] \\
&\le 2\sum_{i=1}^m \psi'(\Gamma_i)\mathbb{E}\big[\|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}}^2\|\delta_{\mu,i}\|_{\nabla^2\phi_i(x_i)^{-1}}^2\big] \\
&\lesssim \|\gamma\|_\infty\|\nabla_\Gamma\Psi(\mu,x)\|_{\tau+\infty}^*\big\|\big(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\delta_{\mu,i}\|_2^2]\big)_{i\in[m]}\big\|_{\tau+\infty} \lesssim \beta^2\epsilon^2\|\nabla_\Gamma\Psi(\mu,x)\|_{\tau+\infty}^*.
\end{aligned}$$

due to (19), Lemma 31, and bound for $\|\gamma\|_\infty$. For the $L_{2,i}^2$ term in (24), we compute

$$\begin{aligned}
\sum_{i=1}^m \psi'(\Gamma_i)\mathbb{E}[L_{2,i}^2] &\lesssim \sum_{i=1}^m \psi'(\Gamma_i)\mathbb{E}\big[\langle \nabla^2\phi_i(x_i)^{\frac{1}{2}}\mathrm{SC}_i^*, \nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\rangle^2\big] \\
&\le \sum_{i=1}^m \psi'(\Gamma_i)\mathbb{E}\big[\|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}}^2\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\|_2^2\big] \\
&\le \|\gamma\|_\infty\|\nabla_\Gamma\Psi(\mu,x)\|_{\tau+\infty}^*\big\|\big(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\|_2^2]\big)_{i\in[m]}\big\|_{\tau+\infty} \lesssim \beta^2\epsilon^2\|\nabla_\Gamma\Psi(\mu,x)\|_{\tau+\infty}^*
\end{aligned}$$

additionally due to Lemma 26.

For the Q-series, we consider the blockwise bounds established in the proof of Lemma 32, and also the stability bounds in Lemma 26 and Lemma 31. It then holds that

$$\sum_{i=1}^m \psi'(\Gamma_i)\sum_{j=1}^5 \mathbb{E}[Q_{j,i}] \lesssim \beta^2\|\nabla_\Gamma\Psi(\mu,x)\|_{\tau+\infty}^*, \qquad \sum_{i=1}^m \psi'(\Gamma_i)\sum_{j=1}^5 \mathbb{E}[Q_{j,i}^2] \lesssim \beta^4\|\nabla_\Gamma\Psi(\mu,x)\|_{\tau+\infty}^*$$

We insert all the above computation back to our first inequality in the proof and obtain

$$\begin{aligned}
\mathbb{E}[\Psi(\mu^{(\mathrm{new})}, x^{(\mathrm{new})})] \le \Psi(\mu,x) + \big\langle \nabla_\mu^{\mathrm{loc}}\Psi(\mu,x), g\big\rangle + (1-\theta/3)\beta\big\|\nabla_\mu^{\mathrm{loc}}\Psi(\mu,x)\big\|_{2,\tau+\infty}^* \\
+ O(\beta^2 + \lambda\beta^2\epsilon^2)\|\nabla_\Gamma\Psi(\mu,x)\|_{\tau+\infty}^*
\end{aligned}$$

additionally due to $\beta \le \epsilon^2$, and $\lambda\|\delta_\Gamma\|_\infty \lesssim \lambda\beta\epsilon \le 1$ from Lemma 32. □

While the descent direction g is computed using the robust approximations $\overline{\mu}$ and $\overline{x}$, the potential change bound in Lemma 34 involves the gradient norm evaluated at the exact values. Therefore, we present two bridging lemmas to bound the difference between the exact and approximate gradient norms.

Lemma 35. *Let $(\overline{x}, \overline{\tau})$ be α-approximation of (x, τ), $\overline{\mu}$ be 10α-approximation of μ, and $\|\Gamma\|_\infty \le \epsilon^2$. If $\|\Gamma - \overline{\Gamma}\|_\infty \le \delta \le \nu \le \min\{\epsilon^2, 0.2\lambda^{-1}\}$ is true, the following property holds*

$$-\frac{\left\langle \nabla_\mu^{\mathrm{loc}} \Psi(\mu, x), g \right\rangle}{\beta} \ge e^{-2\lambda\delta - \alpha} \left(1 - \frac{600\alpha n_{\max}^{\frac{1}{2}}}{\sqrt{\nu}}\right) \|\nabla_\mu^{\mathrm{loc}} \Psi(\mu, x)\|_{2,\tau+\infty}^* - 50\lambda m e^{2\lambda\nu}\sqrt{\nu}$$

Proof. For this proof only, use the shorthand notations

$$\nabla_\mu^{\mathrm{loc}} \Psi = \nabla_\mu^{\mathrm{loc}} \Psi(\mu, x), \qquad \nabla_\mu^{\mathrm{loc}} \overline{\Psi} = \nabla_\mu^{\mathrm{loc}} \Psi(\overline{\mu}, \overline{x}).$$

Also, let us define dual variables

$$a_i = \nabla\phi_i(x_i) - \mu_i, \qquad b_i = \nabla\phi_i(x_i), \qquad \text{and,} \qquad \overline{a}_i = \nabla\phi_i(\overline{x}_i) - \overline{\mu}_i, \qquad \overline{b}_i = \nabla\phi_i(\overline{x}_i).$$

On the basis of the above notations, denote

$$\Delta_{\phi_i}^* = \nabla\phi_i^*(a_i) - \nabla\phi_i^*(b_i), \qquad \overline{\Delta}_{\phi_i}^* = \nabla\phi_i^*(\overline{a}_i) - \nabla\phi_i^*(\overline{b}_i).$$

We first analyze the linear term $\left\langle \nabla_\mu^{\mathrm{loc}} \Psi, g \right\rangle$. From the definition, $\left\langle u_i, (u^\flat)_i \right\rangle \ge 0$ holds for any $u = (u_1, \dots, u_m)$ and $i \in [m]$. As a consequence, the inner product between $\nabla_\mu^{\mathrm{loc}} \Psi$ evaluated at (μ, x) and $(\overline{\mu}, \overline{x})$ can be bounded as

$$\begin{aligned}
\left\langle \nabla_\mu^{\mathrm{loc}} \Psi, \left(\nabla_\mu^{\mathrm{loc}} \overline{\Psi}\right)^{\flat(\overline{\tau})} \right\rangle &= -2 \sum_{i=1}^m \lambda e^{\lambda\Gamma_i} \left\langle \nabla^2\phi_i(x_i)^{\frac{1}{2}} \Delta_{\phi_i}^*, \left(\nabla_\mu^{\mathrm{loc}} \overline{\Psi}\right)_i^{\flat(\overline{\tau})} \right\rangle \\
&= -2 \sum_{i=1}^m \lambda e^{\lambda\Gamma_i} \left\langle \nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}} \overline{\Delta}_{\phi_i}^*, \left(\nabla_\mu^{\mathrm{loc}} \overline{\Psi}\right)_i^{\flat(\overline{\tau})} \right\rangle \\
&\quad -2 \sum_{i=1}^m \lambda e^{\lambda\Gamma_i} \left\langle \nabla^2\phi_i(x_i)^{\frac{1}{2}} \Delta_{\phi_i}^* - \nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}} \overline{\Delta}_{\phi_i}^*, \left(\nabla_\mu^{\mathrm{loc}} \overline{\Psi}\right)_i^{\flat(\overline{\tau})} \right\rangle \\
&= \underbrace{-2 \sum_{i=1}^m \lambda e^{\lambda\Gamma_i} \left\langle \nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}} \overline{\Delta}_{\phi_i}^*, \left(\nabla_\mu^{\mathrm{loc}} \overline{\Psi}\right)_i^{\flat(\overline{\tau})} \right\rangle}_{\text{(I)}} \\
&\quad \underbrace{-2 \sum_{i:\overline{\Gamma}_i \le \nu} \lambda e^{\lambda\Gamma_i} \left\langle \nabla^2\phi_i(x_i)^{\frac{1}{2}} \Delta_{\phi_i}^* - \nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}} \overline{\Delta}_{\phi_i}^*, \left(\nabla_\mu^{\mathrm{loc}} \overline{\Psi}\right)_i^{\flat(\overline{\tau})} \right\rangle}_{\text{(II)}} \\
&\quad \underbrace{-2 \sum_{i:\overline{\Gamma}_i > \nu} \lambda e^{\lambda\Gamma_i} \left\langle \nabla^2\phi_i(x_i)^{\frac{1}{2}} (\Delta_{\phi_i}^* - \overline{\Delta}_{\phi_i}^*), \left(\nabla_\mu^{\mathrm{loc}} \overline{\Psi}\right)_i^{\flat(\overline{\tau})} \right\rangle}_{\text{(III)}}
\end{aligned}$$

$$\underbrace{-2\sum_{i:\overline{\Gamma}_i>\nu}\lambda e^{\lambda\Gamma_i}\left\langle\left(\nabla^2\phi_i(x_i)^{\frac{1}{2}}-\nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}}\right)\overline{\Delta}^*_{\phi_i},\left(\nabla^{\mathrm{loc}}_\mu\overline{\Psi}\right)_i^{\flat(\overline{\tau})}\right\rangle}_{\text{(IV)}}$$

We analyze these terms separately. For the first term, we simply bound

$$-2(\mathrm{I})=\sum_{i=1}^m e^{\lambda(\Gamma_i-\overline{\Gamma}_i)}\left\langle\left(\nabla^{\mathrm{loc}}_\mu\overline{\Psi}\right)_i,\left(\nabla^{\mathrm{loc}}_\mu\overline{\Psi}\right)_i^{\flat(\overline{\tau})}\right\rangle\geq e^{-\lambda\delta}\left\|\nabla^{\mathrm{loc}}_\mu\overline{\Psi}\right\|^*_{2,\overline{\tau}+\infty}.$$

For the second term, it holds that

$$\begin{aligned}|(\mathrm{II})|&\leq\sum_{i:\overline{\Gamma}_i\leq\nu}\lambda e^{\lambda(\nu+\delta)}\left\|\left(\nabla^{\mathrm{loc}}_\mu\overline{\Psi}\right)_i^{\flat(\overline{\tau})}\right\|_2\left(\|\Delta^*_{\phi_i}\|_{\nabla^2\phi_i(x_i)}+\|\overline{\Delta}^*_{\phi_i}\|_{\nabla^2\phi_i(\overline{x}_i)}\right)\\&\leq 2\sum_{i:\overline{\Gamma}_i\leq\nu}\lambda e^{\lambda(\nu+\delta)}\left\|\left(\nabla^{\mathrm{loc}}_\mu\overline{\Psi}\right)_i^{\flat(\overline{\tau})}\right\|_2\left(\|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}}+\|\overline{\mu}_i\|_{\nabla^2\phi_i(\overline{x}_i)^{-1}}\right)\\&\leq 4\lambda e^{\lambda(\nu+\delta)}(\sqrt{\nu+\delta}+\sqrt{\nu})\sum_{i=1}^m\left\|\left(\nabla^{\mathrm{loc}}_\mu\overline{\Psi}\right)_i^{\flat(\overline{\tau})}\right\|_2\leq 12\lambda m e^{2\lambda\nu}\sqrt{\nu}\end{aligned}$$

due to the primal-dual derivative identities from Lemma 5, Lemma 6, $\|\Gamma-\overline{\Gamma}\|_\infty\leq\delta\leq\nu$, $\gamma_i\leq 2\Gamma_i$ (Lemma 8), and $\left\|\left(\nabla^{\mathrm{loc}}_\mu\overline{\Psi}\right)^{\flat(\overline{\tau})}\right\|_{2,\overline{\tau}+\infty}\leq 1$. For the third term, it holds that

$$(\mathrm{III})=\sum_{i:\overline{\Gamma}_i>\nu}e^{\lambda(\Gamma_i-\overline{\Gamma}_i)}\frac{\left(v\right)_i\left(v\right)_i^{\flat(\overline{\tau})}}{2\|\overline{\Delta}^*_{\phi_i}\|_{\nabla^2\phi_i(\overline{x}_i)}}\left\langle\nabla^2\phi_i(x_i)^{\frac{1}{2}}(\overline{\Delta}^*_{\phi_i}-\Delta^*_{\phi_i}),\frac{\nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}}\overline{\Delta}^*_{\phi_i}}{\|\overline{\Delta}^*_{\phi_i}\|_{\nabla^2\phi_i(\overline{x}_i)}}\right\rangle.\tag{26}$$

where we define temporary notation $v_i=2\left(\nabla_\Gamma\overline{\Psi}\right)_i\left\|\overline{\Delta}^*_{\phi_i}\right\|_{\nabla^2\phi_i(\overline{x}_i)}$. In particular, we compute the inner product by repeatedly using primal-dual derivative identities in Lemma 5

$$\begin{aligned}&\left|\left\langle\nabla^2\phi_i(x_i)^{\frac{1}{2}}(\overline{\Delta}^*_{\phi_i}-\Delta^*_{\phi_i}),\frac{\nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}}\overline{\Delta}^*_{\phi_i}}{\|\overline{\Delta}^*_{\phi_i}\|_{\nabla^2\phi_i(\overline{x}_i)}}\right\rangle\right|\\&\leq\|\nabla\phi_i^*(a_i)-\nabla\phi_i^*(\overline{a}_i)\|_{\nabla^2\phi^*(b_i)^{-1}}+\|\nabla\phi_i^*(b_i)-\nabla\phi_i^*(\overline{b}_i)\|_{\nabla^2\phi^*(b_i)^{-1}}\\&\leq 2\|\nabla\phi_i^*(a_i)-\nabla\phi_i^*(\overline{a}_i)\|_{\nabla^2\phi^*(a_i)^{-1}}+\|\nabla\phi_i^*(b_i)-\nabla\phi_i^*(\overline{b}_i)\|_{\nabla^2\phi^*(b_i)^{-1}}\\&\leq 3\|a_i-\overline{a}_i\|_{\nabla^2\phi^*(a_i)}+4\|b_i-\overline{b}_i\|_{\nabla^2\phi^*(b_i)}\\&\leq 4\|\mu_i-\overline{\mu}_i\|_{\nabla^2\phi_i(x_i)^{-1}}+10\|\nabla\phi_i(x_i)-\nabla\phi_i(\overline{x}_i)\|_{\nabla^2\phi_i(x_i)^{-1}}\leq 50\alpha\end{aligned}$$

where, in addition, the first line follows from the Cauchy-Schwarz and the triangular inequality. For the second line we use the following Hessian comparison

$$\begin{aligned}\|\nabla^2\phi_i^*(a_i)\|_{\nabla^2\phi_i^*(b_i)^{-1}}&=\|\nabla^2\phi_i(\nabla\phi_i^*(b_i))\|_{\nabla^2\phi_i(\nabla\phi_i^*(a_i))^{-1}}\\&\leq 1+3\|\nabla\phi_i^*(a_i)-\nabla\phi_i^*(b_i)\|_{\nabla^2\phi_i^*(b_i)^{-1}}\leq 1+6\|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}}\\&\leq 1+12\epsilon\leq 2,\end{aligned}\tag{27}$$

due to self-concordance, Lemma 6, and bound $\gamma_i \le 2\Gamma_i \le 2\epsilon^2$. The third line is due to Lemma 6 again, and the last line uses (27), $\nabla^2\phi_i(x_i) \approx_\alpha \nabla^2\phi_i(\overline{x}_i)$, and duality difference bound from Lemma 33. Besides, we use Lemma 6 for a third time to estimate

$$\|\overline{\Delta}^*_{\phi_i}\|_{\nabla^2\phi_i(\overline{x}_i)} \ge \frac{1}{2}\|\overline{\mu}_i\|_{\nabla^2\phi_i(\overline{x}_i)^{-1}} \ge \frac{1}{4}\sqrt{\overline{\Gamma}_i} \ge \frac{\sqrt{\nu}}{4},$$

additionally also due to Lemma 8. This allows us to compute

$$\begin{aligned}
|(\mathrm{III})| &\le 50\alpha e^{\lambda\delta} \sum_{i:\overline{\Gamma}_i>\nu} \frac{(v)_i (v)_i^{\flat(\overline{\tau})}}{2\|\overline{\Delta}^*_{\phi_i}\|_{\nabla^2\phi_i(\overline{x}_i)}} \\
&= 50\alpha e^{\lambda\delta} \sum_{i:\overline{\Gamma}_i>\nu} \frac{(v)_i (v)_i^{\flat(\overline{\tau})}}{2\|\overline{\Delta}^*_{\phi_i}\|_{\nabla^2\phi_i(\overline{x}_i)}} \frac{\|\overline{\Delta}^*_{\phi_i}\|_{\nabla^2\phi_i(\overline{x}_i)}}{\|\overline{\Delta}^*_{\phi_i}\|_{\nabla^2\phi_i(\overline{x}_i)}} \le \frac{100\alpha e^{\lambda\delta}}{\sqrt{\nu}} \big\|\nabla^{\mathrm{loc}}_\mu \overline{\Psi}\big\|^*_{2,\overline{\tau}+\infty}
\end{aligned}$$

by the definition of the block dual norm. For the final term, it holds that

$$\begin{aligned}
|(\mathrm{IV})| &\le \sum_{i:\overline{\Gamma}_i>\nu} \lambda e^{\lambda\Gamma_i} \left| \left\langle (\nabla^2\phi_i(x_i)^{\frac{1}{2}} - \nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}})\overline{\Delta}^*_{\phi_i}, \big(\nabla^{\mathrm{loc}}_\mu \overline{\Psi}\big)_i^{\flat(\overline{\tau})} \right\rangle \right| \\
&\le e^{\lambda\delta} \sum_{i:\overline{\Gamma}_i>\nu} \lambda e^{\lambda\overline{\Gamma}_i} \big\|\overline{\Delta}^*_{\phi_i}\big\|_{\nabla^2\phi_i(\overline{x}_i)} \big\|\big(\nabla^{\mathrm{loc}}_\mu \overline{\Psi}\big)_i^{\flat(\overline{\tau})}\big\|_2 \big\|\nabla^2\phi_i(\overline{x}_i)^{-\frac{1}{2}}(\nabla^2\phi_i(x_i)^{\frac{1}{2}} - \nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}})\big\|_{\mathrm{op}}
\end{aligned}$$

by Cauchy-Schwarz. In particular, the Hessian comparison term is bounded as

$$\big\|\nabla^2\phi_i(\overline{x}_i)^{-\frac{1}{2}}(\nabla^2\phi_i(x_i)^{\frac{1}{2}} - \nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}})\big\|_{\mathrm{op}} \lesssim \sqrt{n_i}\big\|\nabla^2\phi_i(x_i) - \nabla^2\phi_i(\overline{x}_i)\big\|_{\nabla^2\phi_i(\overline{x}_i)^{-1}} \lesssim \sqrt{n_i}\alpha,$$

using Lemma 22, Lemma 3, and α-approximation. To continue, it holds that

$$|(\mathrm{IV})| \le n_{\max}^{\frac{1}{2}} \alpha e^{\lambda\delta} \sum_{i:\overline{\Gamma}_i>\nu} \big\langle \big(\nabla^{\mathrm{loc}}_\mu \overline{\Psi}\big)_i, \big(\nabla^{\mathrm{loc}}_\mu \overline{\Psi}\big)_i^{\flat(\overline{\tau})} \big\rangle \le n_{\max}^{\frac{1}{2}} \alpha e^{\lambda\delta} \big\|\nabla^{\mathrm{loc}}_\mu \overline{\Psi}\big\|^*_{2,\overline{\tau}+\infty}$$

where we use α-approximation, Lemma 3, and block dual norm definition in the final line. As a result of the above computation, we obtain

$$\begin{aligned}
\left\langle \nabla^{\mathrm{loc}}_\mu \Psi, \big(\nabla^{\mathrm{loc}}_\mu \overline{\Psi}\big)^{\flat(\overline{\tau})} \right\rangle &\ge e^{-\lambda\delta} \left(1 - \alpha e^{2\lambda\delta}\left(\frac{200}{\sqrt{\nu}} + 2n_{\max}^{\frac{1}{2}}\right)\right) \big\|\nabla^{\mathrm{loc}}_\mu \overline{\Psi}\big\|^*_{2,\overline{\tau}+\infty} - 24\lambda m e^{2\lambda\nu}\sqrt{\nu} \\
&\ge e^{-\lambda\delta}\left(1 - \frac{300\alpha n_{\max}^{\frac{1}{2}}}{\sqrt{\nu}}\right) \big\|\nabla^{\mathrm{loc}}_\mu \overline{\Psi}\big\|^*_{2,\overline{\tau}+\infty} - 24\lambda m e^{2\lambda\nu}\sqrt{\nu}
\end{aligned}$$

due to our choice of ν. Now we proceed to the other side. It is noticed that

$$\big\|\big(\nabla^{\mathrm{loc}}_\mu \Psi\big)^{\flat(\tau)}\big\|_{2,\overline{\tau}+\infty} \le e^{\alpha}\big\|\big(\nabla^{\mathrm{loc}}_\mu \Psi\big)^{\flat(\tau)}\big\|_{2,\tau+\infty} \le e^{\alpha},$$

because $\overline{\tau}$ α-approximates τ. Therefore, we reverse the above computation by exchanging the role of $\overline{\Psi}$ and Ψ, with the high/low split now taken according to Γ_i instead of $\overline{\Gamma}_i$, to get

$$\big\|\nabla^{\mathrm{loc}}_\mu \overline{\Psi}\big\|^*_{2,\overline{\tau}+\infty} \ge e^{-\alpha}\left\langle \nabla^{\mathrm{loc}}_\mu \overline{\Psi}, \big(\nabla^{\mathrm{loc}}_\mu \Psi\big)^{\flat(\tau)} \right\rangle$$

$$\geq e^{-\lambda\delta-\alpha}\left(1-\frac{300\alpha n_{\max}^{\frac{1}{2}}}{\sqrt{\nu}}\right)\left\|\nabla_\mu^{\mathrm{loc}}\Psi\right\|_{2,\tau+\infty}^{*}-24\lambda m e^{2\lambda\nu}\sqrt{\nu}$$

where the first step comes from the definition of dual norm and the $\left\|\left(\nabla_\mu^{\mathrm{loc}}\Psi\right)^{\flat(\tau)}\right\|_{2,\overline{\tau}+\infty}\leq e^{\alpha}$. Finally, it holds that

$$\begin{aligned}-\frac{\left\langle\nabla_\mu^{\mathrm{loc}}\Psi,g\right\rangle}{\beta}=\left\langle\nabla_\mu^{\mathrm{loc}}\Psi,\left(\nabla_\mu^{\mathrm{loc}}\overline{\Psi}\right)^{\flat(\overline{\tau})}\right\rangle&\geq e^{-\lambda\delta}\left(1-\frac{300\alpha n_{\max}^{\frac{1}{2}}}{\sqrt{\nu}}\right)\left\|\nabla_\mu^{\mathrm{loc}}\overline{\Psi}\right\|_{2,\overline{\tau}+\infty}^{*}-24\lambda m e^{2\lambda\nu}\sqrt{\nu}\\&\geq e^{-2\lambda\delta-\alpha}\left(1-\frac{600\alpha n_{\max}^{\frac{1}{2}}}{\sqrt{\nu}}\right)\left\|\nabla_\mu^{\mathrm{loc}}\Psi\right\|_{2,\tau+\infty}^{*}-50\lambda m e^{2\lambda\nu}\sqrt{\nu}.\end{aligned}$$

by inserting the definition of g and combining the above computations. □

Lemma 36. *For any $\nu>0$, the following property holds*

$$\left\|\nabla_\Gamma\Psi(\mu,x)\right\|_{\tau+\infty}^{*}\leq\frac{2}{\sqrt{\nu}}\left\|\nabla_\mu^{\mathrm{loc}}\Psi(\mu,x)\right\|_{2,\tau+\infty}^{*}+3\lambda e^{\lambda\nu}m.$$

Proof. For this proof only, use the convenient abbreviation Ψ for $\Psi(\mu,x)$ when it does not cause confusion. The first step is to compute the i-th block of $\nabla_\mu^{\mathrm{loc}}\Psi$

$$(\nabla_\mu^{\mathrm{loc}}\Psi)_i=2\psi'(\Gamma_i)\nabla^2\phi_i(x_i)^{\frac{1}{2}}\nabla_{\mu_i}D_{\phi_i^*}(\nabla\phi_i(x_i)-\mu_i,\nabla\phi_i(x_i))=-2\left(\nabla_\Gamma\Psi\right)_i\nabla^2\phi_i(x_i)^{\frac{1}{2}}\Delta_{\phi_i}^{*}.$$

For any $i\in[m]$, define new block vectors $u,v\in\mathbb{R}^n$ with

$$u_i:=-\nabla^2\phi_i(x_i)^{\frac{1}{2}}\Delta_{\phi_i}^{*},\qquad v_i:=\left(\nabla_\Gamma\Psi(\mu,x)\right)_i^{\flat(\tau)}u_i/\|u_i\|_2.$$

It is easy to verify $\|v_i\|_2=\left(\nabla_\Gamma\Psi(\mu,x)\right)_i^{\flat(\tau)}$. Then the following holds for v

$$\|v\|_{2,\tau+\infty}=\|(\nabla_\Gamma\Psi(\mu,x))^{\flat(\tau)}\|_{\tau+\infty}\leq 1.$$

As a consequence, we utilize the definition of dual norm to obtain

$$\left\|\nabla_\mu^{\mathrm{loc}}\Psi\right\|_{2,\tau+\infty}^{*}=\max_{\|v'\|_{2,\tau+\infty}\leq 1}\left\langle v',\nabla_\mu^{\mathrm{loc}}\Psi\right\rangle\geq\left\langle v,\nabla_\mu^{\mathrm{loc}}\Psi\right\rangle.$$

We compute the lower bound of the inner product

$$\begin{aligned}\left\langle v,\nabla_\mu^{\mathrm{loc}}\Psi\right\rangle&=2\sum_i\left\langle\left(\nabla_\Gamma\Psi\right)_i^{\flat(\tau)}u_i/\|u_i\|_2,\left(\nabla_\Gamma\Psi\right)_i u_i\right\rangle=2\sum_i\|u_i\|_2\left(\nabla_\Gamma\Psi\right)_i\left(\nabla_\Gamma\Psi\right)_i^{\flat(\tau)}\\&=2\sum_{i:\Gamma_i\geq\nu}\|u_i\|_2\left(\nabla_\Gamma\Psi\right)_i\left(\nabla_\Gamma\Psi\right)_i^{\flat(\tau)}+\sum_{i:\Gamma_i<\nu}\left(\nabla_\Gamma\Psi\right)_i\left(\nabla_\Gamma\Psi\right)_i^{\flat(\tau)}\left(2\|u_i\|_2-\sqrt{\nu}+\sqrt{\nu}\right)\\&\geq\frac{\sqrt{\nu}}{2}\sum_i\left(\nabla_\Gamma\Psi\right)_i\left(\nabla_\Gamma\Psi\right)_i^{\flat(\tau)}-\underbrace{\sum_{i:\Gamma_i<\nu}\left(\nabla_\Gamma\Psi\right)_i\left(\nabla_\Gamma\Psi\right)_i^{\flat(\tau)}\left(\sqrt{\nu}-2\|u_i\|_2\right)}_{(*)}\end{aligned}$$

where in the last step we use

$$\|u_i\|_2 = \|\Delta^*_{\phi_i}\|_{\nabla^2\phi_i(x_i)} \geq \frac{1}{2}\|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}} = \frac{1}{2}\sqrt{\gamma_i} \geq \frac{1}{4}\sqrt{\Gamma_i} \geq \frac{\sqrt{\nu}}{4}$$

for any i with $\Gamma_i \geq \nu$. The argument is due to Lemma 6 and Lemma 8. We proceed and compute

$$(*) \leq \sqrt{\nu}\lambda e^{\lambda\nu} \sum_{i:\Gamma_i<\nu} (\nabla_\Gamma \Psi)_i^{\flat(\tau)} \leq \sqrt{\nu}\lambda e^{\lambda\nu} \|(\nabla_\Gamma \Psi)^{\flat(\tau)}\|_1 \leq \sqrt{\nu}\lambda e^{\lambda\nu} m$$

due to the property of ℓ_1-norm. The final result is obtained by combining and rearranging the above inequalities. □

Now we are ready to prove the one-step descent lemma for Algorithm 1.

Proof of Lemma 20. Let $\delta = 100\alpha\epsilon + 800\alpha^2$. By Lemma 33, δ is an upper bound of $\|\Gamma - \overline{\Gamma}\|_\infty$. It is easy to check $\alpha \lesssim \epsilon$ by our parameter choice. Therefore, it follows

$$\delta = 100\alpha\epsilon + 800\alpha^2 \leq 160\alpha\epsilon = \frac{160\theta}{C^2\lambda\sqrt{C_\lambda}} \leq \frac{1}{5\lambda},$$

fulfilling the condition of Lemma 35. Let $\nu := 1/(5\lambda)$. Therefore we invoke Lemma 34 and Lemma 35 to obtain

$$\begin{aligned}
\mathbb{E}[\Psi(\mu^{(\mathrm{new})}, x^{(\mathrm{new})})] - \Psi(\mu, x) \leq & \left\langle \nabla^{\mathrm{loc}}_\mu \Psi(\mu, x), g \right\rangle + (1 - \theta/3)\beta \left\| \nabla^{\mathrm{loc}}_\mu \Psi(\mu, x) \right\|^*_{2,\tau+\infty} \\
& + O(\beta^2 + \lambda\beta^2\epsilon^2) \left\| \nabla_\Gamma \Psi(\mu, x) \right\|^*_{\tau+\infty} + 50\beta\lambda m e^{2\lambda\nu}\sqrt{\nu} \\
\leq & -\beta \left(e^{-2\lambda\delta-\alpha} \left(1 - \frac{600\alpha\sqrt{n_{\max}}}{\sqrt{\nu}} \right) - \left(1 - \frac{\theta}{3} \right) \right) \left\| \nabla^{\mathrm{loc}}_\mu \Psi(\mu, x) \right\|^*_{2,\tau+\infty} \\
& + O(\beta^2 + \lambda\beta^2\epsilon^2) \left\| \nabla_\Gamma \Psi(\mu, x) \right\|^*_{\tau+\infty} + 150\beta m\sqrt{\lambda}.
\end{aligned}$$

We bound the factors before norms. By our choice of parameters in Algorithm 1, it holds that

$$\begin{aligned}
e^{-2\lambda\delta-\alpha} \left(1 - \frac{600\alpha\sqrt{n_{\max}}}{\sqrt{\nu}} \right) - (1 - \theta/3) &\geq \frac{\theta}{3} - \frac{600\alpha\sqrt{n_{\max}}}{\sqrt{\nu}} - 2\lambda\delta - \alpha \\
&\geq \frac{\theta}{3} - \frac{\theta}{12} \geq \frac{\theta}{4}
\end{aligned}$$

because the constants C, C_λ are chosen large enough, and also

$$\beta^2 + \lambda\beta^2\epsilon^2 = \beta^2 (1 + C_\lambda) \lesssim C_\lambda \beta^2.$$

It follows by inserting this back to the descent inequality

$$\mathbb{E}[\Psi(\mu^{(\mathrm{new})}, x^{(\mathrm{new})})] - \Psi(\mu, x) \leq -\frac{\theta\beta}{4} \left\| \nabla^{\mathrm{loc}}_\mu \Psi(\mu, x) \right\|^*_{2,\tau+\infty} + O(C_\lambda\beta^2) \left\| \nabla_\Gamma \Psi(\mu, x) \right\|^*_{\tau+\infty} + \frac{m}{2}$$

additionally due to $150\beta\sqrt{\lambda} = \frac{150\theta}{CC_\lambda} \leq 1/2$. We invoke Lemma 36 with a different $\nu = 1/\lambda$ and compute

$$\mathbb{E}[\Psi(\mu^{(\mathrm{new})}, x^{(\mathrm{new})})] - \Psi(\mu, x) \leq -\frac{\beta\theta\sqrt{\nu}}{8} \left\| \nabla_\Gamma \Psi(\mu, x) \right\|^*_{\tau+\infty} + O(C_\lambda\beta^2) \left\| \nabla_\Gamma \Psi(\mu, x) \right\|^*_{\tau+\infty}$$

$$
\begin{aligned}
&\quad + \frac{3\sqrt{\nu}}{2}\beta\lambda\theta e^{\lambda\nu}m + \frac{m}{2} \\
&\le -\beta\left(\frac{\theta}{8\sqrt{\lambda}} - O\left(\frac{\theta}{C\sqrt{\lambda}}\frac{C_\lambda}{C_\lambda}\right)\right)\left\|\nabla_\Gamma\Psi(\mu,x)\right\|^*_{\tau+\infty} + \frac{3}{2}\frac{em}{10} + \frac{m}{2} \\
&= -\frac{\beta\theta}{\sqrt{\lambda}}\left(\frac{1}{8} - O(C^{-1})\right)\left\|\nabla_\Gamma\Psi(\mu,x)\right\|^*_{\tau+\infty} + m \\
&\le -\frac{\beta\theta}{10\sqrt{\lambda}}\left\|\nabla_\Gamma\Psi(\mu,x)\right\|^*_{\tau+\infty} + m.
\end{aligned}
$$

where in the first line we use $\sqrt{\nu}\beta\lambda\theta = \frac{\theta^2}{CC_\lambda} \le 1/10$. It remains to bound the dual norm of $\nabla_\Gamma\Psi$. We choose $u = \left(2C_{\text{norm}}\sqrt{d}\right)^{-1}\mathbf{1}_m$ and verify

$$
\|u\|_{\tau+\infty} = \frac{1}{2C_{\text{norm}}\sqrt{d}}\|\mathbf{1}\|_{\tau+\infty} \le \frac{C_{\text{norm}}\sqrt{2d}+1}{2C_{\text{norm}}\sqrt{d}} \le 1.
$$

Therefore, by the definition of dual norm, it holds that

$$
\left\|\nabla_\Gamma\Psi(\mu,x)\right\|^*_{\tau+\infty} \ge \sum_{i=1}^m \lambda e^{\lambda\Gamma_i}u_i \ge \frac{\lambda}{2C_{\text{norm}}\sqrt{d}}\sum_{i=1}^m e^{\lambda\Gamma_i} = \frac{\lambda}{2C_{\text{norm}}\sqrt{d}}\Psi(\mu,x).
$$

As a consequence, we obtain

$$
\mathbb{E}[\Psi(\mu^{(\text{new})},x^{(\text{new})})] \le \left(1 - \frac{\beta\theta\sqrt{\lambda}}{20C_{\text{norm}}\sqrt{d}}\right)\Psi(\mu,x) + m \le \left(1 - \frac{\theta^2}{20CC_{\text{norm}}C_\lambda\sqrt{d}}\right)\Psi(\mu,x) + m
$$

by inserting the parameter choices. □

4.5 Feasibility and Sampling Scheme

It remains to establish the approximate primal feasibility still holds after the call to ShortStep, as in the following lemma. The proof is presented in Appendix B.4.

Lemma 37. *Let $(\overline{x},\overline{\tau})$ be an α-approximation of (x,τ). Then w.h.p., the following property holds for $x^{(\text{new})}$ and $\mathbf{H}$ in Algorithm 1*

$$
\left\|\mathbf{A}^\top x^{(\text{new})} - b\right\|_{\mathbf{H}^{-1}} \le \frac{\alpha\beta}{2C_{norm}}.
$$

Also, we provide a valid block-diagonal sampling scheme that meets the requirement of Definition 19. The proof is postponed to Appendix B.5.

Lemma 38 (Sampling scheme)**.** *For $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\delta \in \mathbb{R}^n$, and $\tau \in \mathbb{R}^m_{\ge 0}$, let $p \in \mathbb{R}^m_{\ge 0}$ be defined as*

$$
p_i := \frac{m}{K\sqrt{d}}\left(\frac{\|\delta_i\|_2^2}{\|\delta\|_2^2} + \frac{1}{m}\right) + \frac{\tau_i}{K}
$$

where $K = 2m/\sqrt{d} + \sum_i \tau_i$. Define X to be a random variable that is e_i/p_i with probability p_i, and $R = \frac{1}{N}\sum_{j=1}^N X^{(j)}$, where $X^{(j)}$ are i.i.d. samples of X. If $N \ge 100KC^4_{\text{Var}}\beta^{-2}\log n$, $\mathbf{R} := \mathsf{diag}_{\text{G}}(R)$ is a valid sampling scheme according to Definition 19.

5 Overview of Data Structures

In this section, we introduce the dynamic data structures needed to implement the ideal block Lewis-weight IPM proposed in Section 4. Some of these data structures are block extensions of their counterparts in [vdBLSS20, VDBLL+21, LPT+25]. Among these extensions, some require substantial reimplementation and full proofs, while others follow from simple arguments. We defer all these proofs to Appendix C. In contrast, implementing our ideal IPM also requires a novel finite-sum smoothing data structure. We develop and analyze it in full length in this section.

5.1 Dual and Gradient Maintenance

We first extend the heavy-hitter data structure of [vdBLSS20] to the block setting. Since the blocks have constant size, the same sketching framework and update complexity extend directly to general block-diagonal scaling matrices.

Lemma 39 (Block Heavy Hitter)**.** *There exists a randomized data structure that supports the following operations.*

- *Initialize*$(\mathbf{A} \in \mathbb{R}^{n\times d}, \mathbf{G} \in \mathbb{R}^{n\times n})$ *Takes* $\widetilde{O}(nd+d^2)$ *time to initialize the data structure, where* $\mathbf{G} = \mathsf{diag}_{\mathrm{G}}(\mathbf{G}_1, \ldots, \mathbf{G}_m)$ *should be block-diagonal matrix maintained by the data structure with* $\mathbf{G}_i \in \mathbb{R}^{n_i\times n_i}$ *and* $\sum_i n_i = n$*;*
- *ScaleBlock*$(i, \mathbf{W})$*: sets* $\mathbf{G}_i \leftarrow \mathbf{W}$ *for* $\mathbf{W} \in \mathbb{R}^{n_i,n_i}$ *and takes* $\widetilde{O}(d)$ *time per call;*
- *Query*$(h \in \mathbb{R}^d, \epsilon)$*: w.h.p. outputs a coordinate-level index set containing all* $j \in [n]$ *with* $|(\mathbf{G}\mathbf{A}h)_j| \geq \epsilon$ *in* $\widetilde{O}(\|\mathbf{G}\mathbf{A}h\|_2^2 \cdot d\epsilon^{-2})$ *time per call.*

Algorithm 2: Heavy Hitter Changes

members
  $\mathbf{A} \in \mathbb{R}^{n\times d}$, block diagonal $\mathbf{G} \in \mathbb{R}^{n\times n}$,
  dyadic sketches $\Phi_l \in \mathbb{R}^{m_l\times n}$ and sketched matrices $\mathbf{M}_l \in \mathbb{R}^{m_l\times d}$ for any l,
  JL matrix $\mathbf{R} \in \mathbb{R}^{k\times n}$ and JL sketched matrix $\mathbf{J} \in \mathbb{R}^{k\times d}$
procedure *ScaleBlock*$(i, \mathbf{W})$
  $\mathbf{M}_l \leftarrow \mathbf{M}_l + (\Phi_l)_i(\mathbf{W} - \mathbf{G}_i)\mathbf{A}_i$
  $\mathbf{J} \leftarrow \mathbf{J} + \mathbf{R}_i(\mathbf{W} - \mathbf{G}_i)\mathbf{A}_i$
  $\mathbf{G}_i \leftarrow \mathbf{W}$

Proof. We borrow the HeavyHitter data structure (Algorithm 4) from [vdBLSS20], replacing the procedure Scale with the block analogy ScaleBlock in Algorithm 2, and the member of diagonal matrix g with block-diagonal $\mathbf{G}$. The other procedures remain unchanged in form as Query only returns a coordinate-level heavy hitters. Therefore the correctness, failure probability, and the runtime analysis of Initialize and Query are directly from Lemma 33 in [vdBLSS20]. The new procedure ScaleBlock updates the i-th principal block of $\mathbf{G}$ by adding the sketched difference $\mathbf{W} - \mathbf{G}_i$. The product between $(\Phi_l)_i(\mathbf{W} - \mathbf{G}_i)$ and $\mathbf{A}_i$ takes $\widetilde{O}(d)$ time because $n_i = O(1)$ and Φ_l is $\widetilde{O}(1)$-column sparse. Therefore, the whole procedure ScaleBlock runs in time $\widetilde{O}(d)$. □

We next present a block lazy-update vector-maintenance data structure, which handles the update of dual variable s. The proof is in Appendix C.1.

Lemma 40. *There exists a randomized data structure that works against an adaptive adversary and supports the following procedures*

- *Initialize*$(\mathbf{A} \in \mathbb{R}^{n\times d}, v^{(\text{init})} \in \mathbb{R}^n, \mathbf{W} \in \mathbb{R}^{n\times n}, \epsilon)$*: takes* $\widetilde{O}(nd + d^2)$ *time to initialize the data structure, where* $\mathbf{W} = \mathsf{diag}_{\mathrm{G}}(\mathbf{W}_1, \ldots, \mathbf{W}_m)$ *should be a block-diagonal matrix maintained by the data structure with* $\mathbf{W}_i \in \mathbb{R}^{n_i\times n_i}$ *and* $\sum_i n_i = n$*;*
- *Update*$(i, \mathbf{V})$*: sets* $\mathbf{W}_i \leftarrow \mathbf{V}$ *for* $\mathbf{V} \in \mathbb{R}^{n_i\times n_i}$ *and takes* $\widetilde{O}(d)$ *amortized time per call;*
- *Query*$(h \in \mathbb{R}^d)$*: for the* t*-th call of Query, denote* $v^{(t)} = v^{(\text{init})} + \mathbf{A}\sum_{s=1}^t h^{(s)}$*. Then the* t*-th call outputs w.h.p. a vector* $\overline{v} \in \mathbb{R}^n$ *such that*

$$\left\|\mathbf{W}_i(v_i^{(t)} - \overline{v}_i)\right\|_2 \le \epsilon$$

  holds for any i*. Its total amortized time over* T *calls is*

$$\widetilde{O}\left(Td + \epsilon^{-2}dT\sum_{t=1}^T \left\|\mathbf{W}^{(t)}(v^{(t)} - v^{(t-1)})\right\|_2^2\right);$$

- *ComputeExact*$()$*: outputs the exact value with* $\widetilde{O}(nd)$ *amortized time.*

We present a gradient maintenance structure to output potential descent direction $\nabla_\mu^{\text{loc}}\Psi(\mu, x)$. It combines the bucket strategy in [LS19, vdBLSS20] and the property of block duality vector. The proof is found in Appendix C.2.

Lemma 41 (Gradient maintenance)**.** *There exists a deterministic data structure that supports the following operations*

- *Initialize*$(\mathbf{A} \in \mathbb{R}^{n\times d}, \mu \in \mathbb{R}^n, x \in \mathbb{R}^n, \tau \in \mathbb{R}^m, \epsilon)$*: takes* $\widetilde{O}(nd)$ *time to initialize the data structure. It requires that* $d/n \le \tau \le 2d$*, and* $\|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}} \le \delta$ *for any* $i \in [m]$ *and* $\delta \le 1/10$*;*
- *Update*$(i \in [m], \nu, y, \sigma)$*: sets* $\mu_i \leftarrow \nu$*,* $x_i \leftarrow y$*, and* $\tau_i \leftarrow \sigma$ *for* $\nu \in \mathbb{R}^{n_i}$*,* $y \in \mathbb{R}^{n_i}$*,* $\sigma \in \mathbb{R}$*. It requires* $\|\mu_i\|_{\nabla^2\phi_i(x_i)^{-1}} \le \delta$ *after the update and takes* $\widetilde{O}(d)$ *time per call;*
- *Query*$()$*: outputs* $\nabla_\mu^{\text{loc}}\Psi(\overline{\mu}, x)^{\flat(\overline{\tau})}$ *and* $\mathbf{A}^\top\nabla^2\Phi(x)^{-\frac{1}{2}}\nabla_\mu^{\text{loc}}\Psi(\overline{\mu}, x)^{\flat(\overline{\tau})}$ *for some* $\overline{\tau}$ *and* $\overline{\mu}$ *satisfying* $\overline{\tau} \approx_\epsilon \tau$*, and*

$$\|\overline{\mu}_i - \mu_i\|_{\nabla^2\phi_i(x_i)^{-1}} \le \epsilon, \qquad \text{any } i \in [m],$$

  and takes $\widetilde{O}(n)$ *time.*

5.2 Smoothing Scaling Maintenance

The smoothed preconditioner used in the preceding analysis is defined through an expectation. We now show that it can be efficiently approximated by a finite-sum of independently sampled inverse Hessians. Since each block has constant dimension, Matrix Bernstein concentration guarantees the required spectral approximation with high probability, even under adaptive updates. The resulting data structure only needs to recompute the blocks whose reference points change, without affecting the guarantees of the ideal smoothing.

Algorithm 3: Preconditioner Smoothing

members
  Vector $x \in \mathbb{R}^n$, block diagonal $\mathbf{M} \in \mathbb{R}^{n \times n}$, $\mathbf{S} \in \mathbb{R}^{n \times n}$, and integer K
procedure *Initialize*$(x^{(\text{init})} \in \mathbb{R}^n, \epsilon, \sigma)$
  $x \leftarrow x^{(\text{init})}$
  $\sigma \leftarrow \sigma$
  $\mathbf{M} \leftarrow \nabla^2 \Phi(x)$
  Choose some sufficiently small δ with $\log(1/\delta) = \widetilde{O}(1)$
  $K \leftarrow 1 + 1000\sigma^2 \epsilon^{-2} \log(n_{\max}/\delta)$
  for $i \in [m]$ **do**
    Generate samples $u^{(1)}, \cdots u^{(K)}$ from uniform distribution on the n_i-dimensional unit ball
    $\mathbf{S}_i \leftarrow \frac{1}{K} \sum_{k=1}^K \nabla^2 \phi_i\big(x_i + \sigma \mathbf{M}_i^{-\frac{1}{2}} u^{(k)}\big)^{-1}$
  return S
procedure *Sample*$(i \in [m], y)$
  $x_i \leftarrow y$
  $\mathbf{M}_i \leftarrow \nabla^2 \phi_i(x_i)$
  Generate samples $u^{(1)}, \cdots u^{(K)}$ from uniform distribution on the n_i-dimensional unit ball
  $\mathbf{S}_i \leftarrow \frac{1}{K} \sum_{k=1}^K \nabla^2 \phi_i\big(x_i + \sigma \mathbf{M}_i^{-\frac{1}{2}} u^{(k)}\big)^{-1}$
  return $\mathbf{S}_i$

Lemma 42. *There exists a randomized data structure that works against an adaptive adversary and supports the following procedures:*

- *Initialize*$(x^{(\text{init})}, \epsilon, \sigma)$*: takes* $\widetilde{O}(n\sigma^2/\epsilon^2)$ *time to initialize the data structure and outputs* $\widehat{\mathbf{S}} \in \mathbb{R}^{n \times n}$ *such that w.h.p.* $\widehat{\mathbf{S}}_i \approx_\epsilon \widetilde{\mathbf{S}}_i(x_i, \nabla^2 \phi_i(x_i))$ *for any* $i \in [m]$*;*
- *Sample*(i, y)*: sets* $x_i \leftarrow y$ *and outputs* $\widehat{\mathbf{S}}_i \in \mathbb{R}^{n_i \times n_i}$ *such that w.h.p.* $\widehat{\mathbf{S}}_i \approx_\epsilon \widetilde{\mathbf{S}}_i(x_i, \nabla^2 \phi_i(x_i))$*. It takes* $\widetilde{O}(\sigma^2/\epsilon^2)$ *time per call.*

We need the standard Matrix CLT result from [Tro15] for our analysis.

Proposition 43 (Matrix Bernstein in [Tro15])**.** *Let* $\mathbf{X}_1, ..., \mathbf{X}_n \in \mathbb{R}^{d \times d}$ *be independent, mean-zero, symmetric random matrices such that* $\|\mathbf{X}_i\|_{\text{op}} \le R$ *almost surely for all* $i \in [n]$*. Then for any* $t \ge 0$*,*

$$\Pr\left[\left\|\sum_{i=1}^n \mathbf{X}_i\right\|_{\text{op}} \ge t\right] \le 2d \cdot \exp\left(-\frac{t^2/2}{s^2 + Rt/3}\right)$$

where $s^2 = \left\|\sum_{i=1}^n \mathbb{E}\left[\mathbf{X}_i^2\right]\right\|$*.*

Proof of Lemma 42. Since fresh independent samples are generated at each call, the concentration guarantee holds conditionally on the entire history, and hence also against an adaptive adversary.

Correctness. Fix $i \in [m]$ and omit the subscript. It suffices to show that the finite-sum outputs

$$\widehat{\mathbf{S}} := \frac{1}{K} \sum_{k=1}^K \nabla^2 \phi\left(x + \sigma \mathbf{M}^{-\frac{1}{2}} u^{(k)}\right)^{-1}$$

provides an ϵ-approximation for the ideal smoothed scale $\widetilde{\mathbf{S}} = \mathbb{E}_u \left[\nabla^2\phi\left(x+\sigma\mathbf{M}^{-\frac{1}{2}}u\right)^{-1}\right]$, where $\mathbf{M} = \nabla^2\phi(x)$, and samples $u^{(1)}, \dots u^{(K)}$ are i.i.d. samples from uniform distribution on the unit ball.

Let $\mathbf{S}^{(k)} = \nabla^2\phi\left(x+\sigma\mathbf{M}^{-\frac{1}{2}}u^{(k)}\right)^{-1}$, and $\mathbf{X}^{(k)} = \widetilde{\mathbf{S}}^{-\frac{1}{2}}\mathbf{S}^{(k)}\widetilde{\mathbf{S}}^{-\frac{1}{2}} - \mathbf{I}$. Before we apply the Matrix Bernstein inequality, it requires to examine the assumptions. First, for u in the unit ball, it holds that

$$\nabla^2\phi(x+\sigma\mathbf{M}^{-\frac{1}{2}}u^{(k)})^{-1} \preceq \frac{1}{(1-\sigma)^4}\nabla^2\phi(x+\sigma\mathbf{M}^{-\frac{1}{2}}u)^{-1} \preceq e^{6\sigma}\nabla^2\phi(x+\sigma\mathbf{M}^{-\frac{1}{2}}u)^{-1}.$$

by using the Hessian comparison in Lemma 10. A corresponding lower bound can be obtained similarly. Taking expectation over u immediately suggests $\mathbf{S}^{(k)} \approx_{6\sigma} \widetilde{\mathbf{S}}$. Therefore $\left\|\mathbf{X}^{(k)}\right\|_{\mathrm{op}} \le 10\sigma$ by our choice of σ. Moreover, we compute

$$\left\|\sum_{k=1}^{K}\mathbb{E}\left[\left(\mathbf{X}^{(k)}\right)^2\right]\right\|_{\mathrm{op}} \le K\left\|\mathbb{E}\left[\left(\mathbf{X}^{(1)}\right)^2\right]\right\|_{\mathrm{op}} \le K\mathbb{E}\left[\left\|\mathbf{X}^{(1)}\right\|_{\mathrm{op}}^2\right] \le 100K\sigma^2.$$

Finally, it is easy to assert that $\mathbf{X}^{(k)}$ is zero-mean and i.i.d. from its definition. Now let $K = 1 + 1000\sigma^2\epsilon^{-2}\log(n_{\max}/\delta)$, $t = \epsilon K$. We invoke Proposition 43 to obtain

$$\Pr\left[\left\|\widetilde{\mathbf{S}}^{-\frac{1}{2}}(\widehat{\mathbf{S}}-\widetilde{\mathbf{S}})\widetilde{\mathbf{S}}^{-\frac{1}{2}}\right\|_{\mathrm{op}} \ge \epsilon\right] = \Pr\left[\left\|\sum_{k=1}^{K}\mathbf{X}^{(k)}\right\|_{\mathrm{op}} \ge t\right] \le \delta.$$

This immediately suggests that w.h.p. $\widehat{\mathbf{S}} \approx_\epsilon \widetilde{\mathbf{S}}$.

Time analysis. This is immediate as computing $\mathbf{S}_i$ for any block $i \in [m]$ takes $K = \widetilde{O}(\sigma^2/\epsilon^2)$ samples and $O(n_i) = 1$. □

5.3 Lewis Weight Maintenance

For Lewis-weight, we first extend the leverage-score maintenance data structure of [VDBLL+21] to the block setting. Since the Lewis-weight maintenance algorithm in [VDBLL+21] uses leverage-score maintenance as a backend, our block version can be incorporated into the same framework. Lemma 45 ensures that its soundness.

Lemma 44 (Leverage score)**.** *There exists a randomized data structure that works against an adaptive adversary. Suppose the following conditions are true*

- *There exists a sequence $\widetilde{\mathbf{V}}^{(t)}$ such that, $(\mathbf{V}^{(t)})^2 \approx_{0.02/\log T} (\widetilde{\mathbf{V}}^{(t)})^2$, $(\mathbf{V}^{(t)})^2 \approx_{0.02/\log T} (\mathbf{V}^{(t-1)})^2$ and*
$$\left\|\left(\left\|(\widetilde{\mathbf{V}}_i^{(t-1)})^{-1}(\widetilde{\mathbf{V}}_i^{(t)} - \widetilde{\mathbf{V}}_i^{(t-1)})\right\|_{\mathrm{op}}\right)_{i\in[m]}\right\|_{\sigma(\widetilde{\mathbf{V}}^{(t-1)}\mathbf{A})} \le O(1)$$
are true for all $t \in [T]$.
- *There exists a black box solver that takes $T_{inverse}$ time per call to solve the weighted linear system $\mathbf{A}^\top\mathbf{B}\mathbf{A}x = b$ for some block-diagonal $\mathbf{B} \succ 0$ and vector b up to some spectral precision.*

The data structure supports the following properties

- *Initialize*($\mathbf{A} \in \mathbb{R}^{n\times d}, \mathbf{V} \in \mathbb{R}^{n\times n}, \epsilon$): *takes* $\widetilde{O}(\epsilon^{-2}nd + \epsilon^{-2}T_{inverse} + d^2)$ *time to initialize the data structure, where* $\mathbf{V} = \mathsf{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_m)$ *should be a block-diagonal matrix maintained by the data structure with* $\mathbf{V}_i \in \mathbb{R}^{n_i\times n_i}$ *and* $\sum_i n_i = n$*;*
- *Update*($i, \mathbf{W}$): *sets* $\mathbf{V}_i \leftarrow \mathbf{W}$ *and takes* $\widetilde{O}(d\epsilon^{-2} + m\epsilon^{-4}\sigma(\mathbf{VA})_i)$ *amortized time per call;*
- *Query*(): *for the* t*-th call of Query, denote* $\mathbf{V}^{(t)}$ *to be the scaling matrix before the call. Then the* t*-th call outputs w.h.p. a vector* $\overline{\sigma} \in \mathbb{R}^m$ *such that* $\overline{\sigma} \approx_\epsilon \sigma(\mathbf{V}^{(t)}\mathbf{A}) + d/m\mathbf{1}_m$*. Its amortized time is* $\widetilde{O}(d^2\epsilon^{-4} + Tm\epsilon^{-4} + \epsilon^{-2}T_{inverse})$ *per call.*

The proof of Lemma 44 is postponed to Appendix C.3.

Lemma 45. *Consider* w*,* z *and* $\overline{\sigma}$ *with* $w \approx_\epsilon \sigma(\mathbf{W}^{\frac{1}{2}-\frac{1}{p}}\mathbf{A}) + z$*,* $\overline{\sigma} \approx_\gamma \sigma(\mathbf{W}^{\frac{1}{2}-\frac{1}{p}}\mathbf{A}) + z$*, and* $p \in (0, 2)$*. Define* $w' = (w^{2/p-1}\overline{\sigma})^{p/2}$*, then* $w' \approx_{(\epsilon+\gamma)|1-p/2|+\gamma} \sigma((\mathbf{W}')^{\frac{1}{2}-\frac{1}{p}}\mathbf{A}) + z$*,* $w' \approx_{(\epsilon+\gamma)p/2} w$

Proof. The lemma is a block version of Lemma 5.3 in [VDBLL+21]. Since $w \approx_{\gamma+\epsilon} \overline{\sigma}$, it follows that

$$w' = (w^{2/p-1}\overline{\sigma})^{p/2} = w \cdot \left(\frac{\overline{\sigma}}{w}\right)^{p/2} \approx_{(\epsilon+\gamma)p/2} w.$$

For the second claim, the same scalar calculation as in Lemma 5.3 of [VDBLL+21] and the definition of block leverage score gives

$$\frac{\sigma(\mathbf{W}'^{\frac{1}{2}-\frac{1}{p}}\mathbf{A})_i + z_i}{w'_i} \approx_\gamma \frac{\sum_{j\in \mathrm{G}_i} a_j^\top (\mathbf{A}^\top (\mathbf{W}'_{\mathrm{G}})^{1-2/p}\mathbf{A})^{-1} a_j + (w'_i)^{2/p-1} z_i}{\sum_{j\in \mathrm{G}_i} a_j^\top (\mathbf{A}^\top \mathbf{W}_{\mathrm{G}}^{1-2/p}\mathbf{A})^{-1} a_j + w_i^{2/p-1} z_i}.$$

By the first claim, $\mathbf{W}_{\mathrm{G}} = \mathsf{diag}_{\mathrm{G}}(w) \approx_{\epsilon+\gamma} \mathbf{W}_{\mathrm{G}} = \mathsf{diag}_{\mathrm{G}}(w') = \mathbf{W}'_{\mathrm{G}}$. Consequently it holds that $w' \approx_{(\epsilon+\gamma)|1-p/2|+\gamma} \sigma(\mathbf{W}'^{\frac{1}{2}-\frac{1}{p}}\mathbf{A}) + z$. □

Lemma 46 (Lewis weight maintenance)**.** *There exists a randomized data structure that works against an adaptive adversary. Suppose the following conditions hold:*

- *The stability conditions of Lemma 44 hold for the leverage-score data structures used in the Lewis-weight maintenance. Moreover, the scaling changes satisfy* $(\mathbf{V}_i^{(t)})^2 \approx_{O(\delta\epsilon)} (\mathbf{V}_i^{(t-1)})^2$ *for every* $i \in [m]$ *and* $t \in [T]$*, where* $\delta > 1$ *and* $\epsilon \lesssim 1/(\delta \log m)$*.*
- *There exists a black-box solver satisfying the same conditions as in Lemma 44, with running time* $\widetilde{O}(T_{inverse})$ *per inverse action.*

The data structure supports the following operations:

- *Initialize*($\mathbf{A} \in \mathbb{R}^{n\times d}, \mathbf{V} \in \mathbb{R}^{n\times n}, p, \epsilon, \delta$): *takes* $\widetilde{O}(\epsilon^{-2}nd + \epsilon^{-2}T_{inverse} + d^2)$ *time to initialize the data structure;*
- *Scale*($i, \mathbf{W}$): *sets* $\mathbf{V}_i \leftarrow \mathbf{W}$ *for a positive definite matrix* $\mathbf{W} \in \mathbb{R}^{n_i\times n_i}$ *and takes amortized time* $\widetilde{O}(m\epsilon^{-O(\log\delta)}w_p(\mathbf{VA})_i)$ *per call;*
- *Query*(): *for the* t*-th call, let* $\mathbf{V}^{(t)}$ *denote the scaling matrix before the call. The procedure outputs w.h.p. a vector* $\overline{\tau} \in \mathbb{R}^m$ *satisfying* $\overline{\tau} \approx_\epsilon \tau(\mathbf{V}^{(t)}\mathbf{A})$*, together with the set of blocks whose estimates have changed. Its amortized time is* $\widetilde{O}(d^2\epsilon^{-4} + Tm\epsilon^{-6} + \epsilon^{-2}T_{inverse})$ *per call.*

Proof. We use the Algorithm 3 in [VDBLL+21], replacing the scalar leverage score maintenance described by its Theorem C.1 and C.2 with the newly implemented block version in our Lemma 44. Because Lemma 45 provides the approximate iteration for block Lewis, and the Lemma 44 subsumes the same condition and output, it is immediate that the correct and time analysis in Section 5 also holds for our setting, as all the analysis only relies on leverage score as a backend and does not rely on the block structure. □

5.4 Sampling and Inverse Maintenance

We provide a block sampling maintenance data structure, almost directly adapted from Lemma 6.3 in [LPT+25] [vdBLSS20]. The proof is postponed to Appendix C.4.

Lemma 47 (Sampling Maintenance)**.** *There exists a randomized data structure that works against adaptive queries and supports the following procedures:*

- *Initialize*$(\mathbf{A} \in \mathbb{R}^{n\times d}, \mathbf{W} \in \mathbb{R}^{n\times n}, \tau \in \mathbb{R}^m_{\geq 0})$*: takes* $\widetilde{O}(nd)$ *time to initialize the data structure, where* $\mathbf{W} = \mathsf{diag}_{\mathrm{G}}(\mathbf{W}_1, \ldots, \mathbf{W}_m)$ *is block-diagonal with* $\mathbf{W}_i \in \mathbb{R}^{n_i \times n_i}$ *and* $\sum_{i=1}^m n_i = n$*;*
- *Update*$(i, \mathbf{V}, \mu)$*: sets* $\mathbf{W}_i \leftarrow \mathbf{V}$ *for* $\mathbf{V} \in \mathbb{R}^{n_i \times n_i}$ *and* $\tau_i \leftarrow \mu$ *for* $\mu \geq 0$*, and takes amortized time* $\widetilde{O}(d)$ *per call;*
- *Sample*$(h \in \mathbb{R}^d)$*: with* $\delta = \mathbf{W}\mathbf{A}h$ *and* $K = 2m/\sqrt{d} + \sum_i \tau_i$*, independently samples* N *blocks according to*

$$p_i = \frac{m}{K\sqrt{d}} \left(\frac{\|\delta_i\|_2^2}{\|\delta\|_2^2} + \frac{1}{m} \right) + \frac{\tau_i}{K},$$

and outputs w.h.p. a valid sampling matrix $\mathbf{R} = \mathsf{diag}_{\mathrm{G}}(R)$*, where*

$$R_i = \frac{\#\{j : i_j = i\}}{Np_i}.$$

It takes $\widetilde{O}(Nd)$ *amortized time per call.*

The following lemma extends the inverse maintenance in [vdBLSS20] to the block version. The proof is postponed to Appendix C.4.

Lemma 48 (Block Inverse Maintenance)**.** *There exists a randomized data structure that works against an adaptive adversary. Suppose the following conditions are true:*

- *Let* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *have full column rank. Let* $\mathbf{W} = \mathsf{diag}_{\mathrm{G}}(\mathbf{W}_1, \ldots, \mathbf{W}_m)$ *be a block-diagonal positive definite matrix with* $\mathbf{W}_i \in \mathbb{R}^{n_i \times n_i}$ *satisfying* $n_i = O(1)$ *and* $\sum_i n_i = n$*.. Define*

$$\mathbf{K}(\mathbf{W}) := \mathbf{A}^\top \mathbf{W} \mathbf{A}, \qquad l_i(\mathbf{W}) := \mathrm{tr}\left(\mathbf{K}(\mathbf{W})^{-1} \mathbf{A}_i^\top \mathbf{W}_i \mathbf{A}_i \right).$$

The leverage hints satisfy $\tau_i \geq l_i(\mathbf{W})$ *for all* $i \in [m]$ *and* $\sum_i \tau_i = O(d)$*.*

- *Let* $(\mathbf{W}^{(t)}, \tau^{(t)})_{t=0}^T$ *denote the weights and leverage hints before successive calls to Solve. There exist reference sequences* $(\widehat{\mathbf{W}}^{(t)}, \widehat{\tau}^{(t)})_{t=0}^T$ *such that*

$$\left\| \mathbf{W}_i^{(t)} - \widehat{\mathbf{W}}_i^{(t)} \right\|_{(\widehat{\mathbf{W}}_i^{(t)})^{-1}} \leq \frac{\epsilon}{cL}, \qquad \tau_i^{(t)} \approx_{1/(cL)} \widehat{\tau}_i^{(t)}$$

for all $i \in [m]$ and $t \in \{0, \ldots, T\}$, where c is a sufficiently large constant and $L = \Theta(\log d)$. Furthermore,

$$\left\| \left(\epsilon^{-1} \left\| \widehat{\mathbf{W}}_i^{(t)} - \widehat{\mathbf{W}}_i^{(t-1)} \right\|_{(\widehat{\mathbf{W}}_i^{(t-1)})^{-1}} + \left| (\widehat{\tau}_i^{(t)} - \widehat{\tau}_i^{(t-1)}) / \widehat{\tau}_i^{(t-1)} \right| \right)_{i \in [m]} \right\|_{\max\{\widehat{\tau}^{(t)}, \widehat{\tau}^{(t-1)}\}} \leq O(1)$$

for all $t \in [T]$, where the maximum is taken coordinatewise. The reference sequences need not be computed.

The data structure supports the following operations:

- *INITIALIZE$(\mathbf{A}, \mathbf{W}, \tau, \epsilon)$: takes $\widetilde{O}(nd + \epsilon^{-2} d^{\omega})$ time to initialize the data structure, where $\epsilon \in (0, 1/10)$ is the target precision;*
- *UPDATE$(i, \mathbf{V}, \mu)$: sets $\mathbf{W}_i \leftarrow \mathbf{V}$ for a positive definite matrix $\mathbf{V} \in \mathbb{R}^{n_i \times n_i}$ and $\tau_i \leftarrow \mu$ for $\mu \geq 0$, and takes $\widetilde{O}(1)$ amortized time per call;*
- *SOLVE$(b, \mathbf{W}')$: given $b \in \mathbb{R}^d$ and a block-diagonal positive definite matrix $\mathbf{W}'$ satisfying $\mathbf{W}'_i \approx_1 \mathbf{W}_i$ for all $i \in [m]$, outputs w.h.p. a vector $y \in \mathbb{R}^d$ such that*

$$y = \mathbf{H}^{-1} b \qquad \text{for some} \qquad \mathbf{H} \approx_{\epsilon} \mathbf{K}(\mathbf{W}').$$

Its amortized running time is $\widetilde{O}(n + \epsilon^{-2}(d^2 + d^{\omega - 1/2}))$ per call.

6 Implementation of Path Following Algorithm

In this section, we show how to efficiently implement the ideal Lewis-weight IPM given by Algorithm 1. We will analyze the total runtime of the implementation, which takes $\widetilde{O}(\sqrt{d})$ iteration to converge by previous analysis.

6.1 Initial and Final Processing

It requires the following initial and final processing procedures. In particular, we need to the efficient computation of a proper initial point with a extended program. Also, it requires to restore strict primal feasibility and reduces to the original program after $\widetilde{O}(\sqrt{d})$ iterations. The proofs of the following lemmas are postponed to Appendix D.1.

Lemma 49. *Consider the programming in (2) with an accuracy parameter $\delta > 0$. Suppose that $\|x\|_{2,\infty} \leq R$ for any $x \in K$, $\|c\|_{2,\infty} \leq L$, and $\max_{i \in [m]} \|\mathbf{A}_i\|_{\mathrm{op}} \leq M$. Its extended program*

$$\min_{\substack{\widetilde{x} \in K \times [0,2] \\ \widetilde{\mathbf{A}}^{\top} \widetilde{x} = \widetilde{b}}} \widetilde{c}^{\top} \widetilde{x} \tag{28}$$

with

$$\widetilde{\mathbf{A}} = \begin{bmatrix} \mathbf{A} \\ b - \mathbf{A}^{\top} x^{(\mathrm{init})} \end{bmatrix}, \qquad \widetilde{b} = b, \qquad \text{and} \qquad \widetilde{c} = \begin{bmatrix} c \\ \frac{LR}{\delta} \end{bmatrix}$$

satisfy the following properties:

- *Let* $x^{(\text{init})} = \operatorname{argmin}_x \sum_{i=1}^m \phi_i(x_i)$, $s^{(\text{init})} = c$, $z^{(\text{init})} = 1$, $r^{(\text{init})} = LR/\delta$, *and* $\rho = 5RmL/(d\delta^2)$. *The triplet* $\left(\begin{bmatrix} x^{(\text{init})} \\ z^{(\text{init})} \end{bmatrix}, \begin{bmatrix} s^{(\text{init})} \\ r^{(\text{init})} \end{bmatrix}, \rho^{(\text{init})}\right)$ *is* δ*-centered for barrier* $\widetilde{\Phi}([x,z]) = \sum_{i=1}^m \phi_i(x_i) - h(z)$ *with* $h(z) = \log(z) + \log(2-z)$;
- *For any* $\widetilde{x} = \begin{bmatrix} x \\ z \end{bmatrix}$ *such that* $\widetilde{\mathbf{A}}^\top \widetilde{x} = \widetilde{b}$, $\overline{x} \in K \times [0,2]$, *and* $\widetilde{c}^\top \widetilde{x} \le \min_{\widetilde{\mathbf{A}}^\top \widetilde{x} = \widetilde{b}, \widetilde{x} \in K \times [0,2]} \widetilde{c}^\top \widetilde{x} + \delta$, *then* x *satisfies*

$$c^\top x \le \min_{x \in K, \mathbf{A}^\top x = b} c^\top x + \delta, \qquad \left\|\mathbf{A}^\top x - b\right\|_\infty \le 2mRM\left(2m\delta + \frac{\delta^2}{LR}\right).$$

Lemma 50. *For an* ϵ*-centered point* (x, s, ρ) *for* (2)*, we can compute a feasible pair* (x, s) *such that*

$$c^\top x - \min_{x \in K, \mathbf{A}^\top x = b} c^\top x \le 8\rho d$$

6.2 Nearby Sequence Properties

Before describing the path-following implementation and time analysis, it requires to establish the following nearby sequence properties. Their proofs are postponed to Appendix D.2.

Lemma 51. *Let* $(x^{(k)}, s^{(k)})$ *for* $k \in [T]$ *be the sequence of points generated by repeatedly calling Algorithm 1. For any fixed* $\alpha_{\text{nb}} \in (0, \beta]$*, if* $C_{\text{Var}}^2 \ge \beta/\alpha_{\text{nb}}$ *holds, the following properties hold w.h.p.*

- $\left\|\nabla^2\Phi(x^{(k)})^{\frac{1}{2}}(\widehat{x}^{(k)} - x^{(k)})\right\|_{2,\infty} \le \alpha_{\text{nb}}$;
- $\left\|\nabla^2\Phi(\widehat{x}^{(k)})^{\frac{1}{2}}(\widehat{x}^{(k)} - x^{(k)})\right\|_{2,\infty} \le 2\alpha_{\text{nb}}$;
- $\left\|\nabla^2\Phi(\widehat{x}^{(k)})^{\frac{1}{2}}(\widehat{x}^{(k+1)} - \widehat{x}^{(k)})\right\|_{2,\tau(x^{(k)})+\infty} \le 2\beta$

Lemma 52. *Suppose the setting of Lemma 51. Define* $\mathbf{C}^{(k)} = \widetilde{\mathbf{C}}(x^{(k)})$, $\widehat{\mathbf{C}}^{(k)} = \widetilde{\mathbf{C}}(\widehat{x}^{(k)})$, $\tau^{(k)} = \tau(x^{(k)})$, *and* $\widehat{\tau}^{(k)} = \tau(\widehat{x}^{(k)})$. *Let* $\mathbf{W}^{(k)} = \mathbf{T}(x^{(k)})^{1/2-1/p}\mathbf{C}^{(k)}$ *and* $\widehat{\mathbf{W}}^{(k)} = \mathbf{T}(\widehat{x}^{(k)})^{1/2-1/p}\widehat{\mathbf{C}}^{(k)}$. *Then the following properties hold w.h.p.:*

- $\max_{i\in[m]} \left\|(\widehat{\mathbf{W}}_i^{(k)})^{-1}(\mathbf{W}_i^{(k)} - \widehat{\mathbf{W}}_i^{(k)})\right\|_{\text{op}} \lesssim \alpha_{\text{nb}}$;
- $\left\|\mathbf{T}(\widehat{x}^{(k)})^{-1}(\widehat{\tau}^{(k+1)} - \widehat{\tau}^{(k)})\right\|_{\widehat{\tau}^{(k)}} \lesssim \frac{\beta}{C_{\text{norm}}}$;
- $\left\|\left(\left\|(\widehat{\mathbf{W}}_i^{(k)})^{-1}(\widehat{\mathbf{W}}_i^{(k+1)} - \widehat{\mathbf{W}}_i^{(k)})\right\|_{\text{op}}\right)_{i\in[m]}\right\|_{\widehat{\tau}^{(k)}} \lesssim \frac{\beta}{C_{\text{norm}}}$.

6.3 Short-Step Implementation

We present the implementation of the ideal ShortStep in Algorithm 5, and also the entire path-following algorithm in Algorithm 4. We establish the correctness in this subsection, and will discuss the running cost in the next subsection.

Lemma 53. *Algorithm 5 correctly implements Short-Step in Algorithm 1.*

Algorithm 4: Block Path Following

procedure BLOCKPATHFOLLOWING($\mathbf{A}, x^{(\text{init})}, s^{(\text{init})}, \rho^{(\text{init})}, \rho^{(\text{fi})}$)
  Set parameters $p, \alpha, \beta, \epsilon, \lambda, \sigma, r$ as in Algorithm 1
  $x \leftarrow x^{(\text{init})}$, $\overline{x} \leftarrow x^{(\text{init})}$, $\overline{s} \leftarrow s^{(\text{init})}$, $\overline{\rho} \leftarrow \rho^{(\text{init})}$, $\rho \leftarrow \rho^{(\text{init})}$
  $h_2 \leftarrow \mathbf{A}^\top x - b$
  $\overline{\mathbf{S}} \leftarrow D^{(\text{smooth})}.\text{INITIALIZE}(x, \alpha/2^{24}, \sigma)$
  $\overline{\tau} \leftarrow D^{(\tau)}.\text{INITIALIZE}(\mathbf{A}, \overline{\mathbf{S}}^{\frac{1}{2}}, p, \alpha/2^{24}, C_\tau)$
  $D^{(s)}.\text{INITIALIZE}(\mathbf{A}, \overline{s}, \rho^{-1}\overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}, \alpha/2^{24})$
  $D^{(\text{sample})}.\text{INITIALIZE}(\mathbf{A}, \overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}, \overline{\tau})$
  $D^{(-1)}.\text{INITIALIZE}(\mathbf{A}, \overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-1}, \overline{\tau}, \alpha/2^3)$
  $D^{(\nabla)}.\text{INITIALIZE}(\mathbf{A}, \overline{\mathbf{T}}^{-1}\overline{s}/\rho + \nabla\Phi(x), \overline{x}, \overline{\tau}, \alpha/2^{24})$
  while $\rho > \rho^{(\text{fi})}$ **do**
    Execute SHORTSTEPIMPLEMENT(ρ)
    $\rho \leftarrow (1-r)\rho$
  $s \leftarrow D^{(s)}.\text{COMPUTEEXACT}()$
  return x, s

Proof. We first introduce the data structures used in Algorithm 5. We use $D^{(s)}$, $D^{(\nabla)}$, $D^{(\tau)}$, $D^{(\text{smooth})}$, $D^{(\text{sample})}$, and $D^{(-1)}$ to denote the data structures in Lemma 40, Lemma 41, Lemma 46, Lemma 42, Lemma 47, and Lemma 48, respectively.

While the ideal algorithm uses approximate variables $(\overline{x}, \overline{\tau}, \overline{\mu})$ for (x, τ, μ), Algorithm 5 instead explicitly maintains a different $(\overline{x}, \overline{s}, \overline{\tau})$. The correspondence for $\overline{x}$ and $\overline{\tau}$ is direct as they represent the corresponding approximate variables in Algorithm 1. In contrast, the residual approximation $\overline{\mu}$ is not maintained explicitly. More precisely, for every call to $D^{(\nabla)}.\text{QUERY}$ and the output (h_g, h_1), there exists some $\overline{\mu}$ such that

$$h_g = -\beta\nabla_\mu^{\text{loc}}(\overline{\mu}, \overline{x})^{\flat(\overline{\tau})}, \quad h_1 = -\beta\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\nabla_\mu^{\text{loc}}(\overline{\mu}, \overline{x})^{\flat(\overline{\tau})}.$$

We will verify that $\overline{\mu}$ satisfies the required approximation guarantee, and the maintained approximations $(\overline{x}, \overline{s}, \overline{\tau})$ remain accurate after each call to SHORTSTEP.

We prove by induction that every call to SHORTSTEPIMPLEMENT makes a valid call to SHORTSTEP in Algorithm 1. Suppose that the current variables satisfy the inductive guarantees of Algorithm 1. Also, suppose immediately before the call, the maintained variables $(\overline{x}, \overline{s}, \overline{\tau})$ satisfy, for every $i \in [m]$,

$$\left\|\nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}}(\overline{x}_i - x_i)\right\|_2 \le \frac{\alpha}{2^{10}}, \qquad \frac{1}{\overline{\rho\tau}_i}\left\|\nabla^2\phi_i(\overline{x}_i)^{-\frac{1}{2}}(\overline{s}_i - s_i)\right\|_2 \le \frac{\alpha}{2^{10}}, \qquad \overline{\tau}_i \approx_{\alpha/2^{10}} \tau_i(x) \tag{29}$$

We first show $\overline{\mu}$ implicitly maintained by $D^{(\nabla)}$ is an $\alpha/2^5$-approximation of the current μ. Let $(\overline{x}^{(\text{new})}, \overline{s}^{(\text{new})}, \overline{\tau}^{(\text{new})})$ denote the maintained variables after the call. We need to show that the bounds in (29) still hold. By our choice of parameters, these bounds imply the approximation invariance in Definition 18.

Algorithm 5: Implementation of ShortStep

procedure *ShortStepImplement*(ρ)
  if $\overline{\rho} \not\approx_{\alpha/2^{14}} \rho^{(\text{new})}$ **then**
    $\overline{\rho} \leftarrow \rho^{(\text{new})}$
    for $i \in [m]$ **do**
      $D^{(\nabla)}.\textsc{Update}(i, \overline{s}_i/(\overline{\rho\tau}_i) + \nabla\phi_i(\overline{x}_i), \overline{x}_i, \overline{\tau}_i)$
      $D^{(s)}.\textsc{Update}(i, \nabla^2\phi_i(\overline{x}_i)^{-\frac{1}{2}}/(\overline{\rho\tau}_i))$
  $(h_g, h_1) \leftarrow -\beta D^{(\nabla)}.\textsc{Query}()$
  $h_{\text{root}} \leftarrow D^{(-1)}.\textsc{Solve}(h_1 + h_2, \overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-1})$
  $h_s \leftarrow D^{(-1)}.\textsc{Solve}(h_1, \overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-1})$
  $\mathbf{R} \leftarrow D^{(\text{sample})}.\textsc{Sample}(h_{\text{root}})$
  $h_r = \overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}h_{\text{root}}$, $h_2 \leftarrow h_2 + h_1 - \mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{R}h_r$
  $x \leftarrow x + \nabla^2\Phi(\overline{x})^{-\frac{1}{2}}h_g - \nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{R}h_r$
  for $i \in [m]$ **do**
    if $\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}(x_i - \overline{x}_i)\right\|_2 \geq \alpha/2^{20}$ **then**
      $\overline{x}_i \leftarrow x_i$
      $D^{(\nabla)}.\textsc{Update}(i, \overline{s}_i/(\overline{\rho\tau}_i) + \nabla\phi_i(\overline{x}_i), \overline{x}_i, \overline{\tau}_i)$
      $D^{(\tau)}.\textsc{Scale}(i, D^{(\text{smooth})}.\textsc{Sample}(i, \overline{x}_i)^{\frac{1}{2}})$
      $D^{(\text{sample})}.\textsc{Scale}(i, \overline{\tau}_i^{-1}\nabla^2\phi_i(\overline{x}_i)^{-\frac{1}{2}}, \overline{\tau}_i)$
      $D^{(-1)}.\textsc{Update}(i, \overline{\tau}_i^{-1}\nabla^2\phi_i(\overline{x}_i)^{-1}, \overline{\tau}_i)$
      $D^{(s)}.\textsc{Update}(i, \nabla^2\phi_i(\overline{x}_i)^{-\frac{1}{2}}/(\overline{\rho\tau}_i))$
  $\tau^{\text{tmp}}, I_\tau = D^{(\tau)}.\textsc{Query}()$
  for $i \in I_\tau$ **do**
    if $\tau_i^{\text{tmp}} \not\approx_{\alpha/2^{12}} \overline{\tau}_i$ **then**
      $\overline{\tau}_i \leftarrow \tau_i^{\text{tmp}}$
      $D^{(\nabla)}.\textsc{Update}(i, \overline{s}_i/(\overline{\rho\tau}_i) + \nabla\phi_i(\overline{x}_i), \overline{x}_i, \overline{\tau}_i)$
      $D^{(\text{sample})}.\textsc{Scale}(i, \overline{\tau}_i^{-1}\nabla^2\phi_i(\overline{x}_i)^{-\frac{1}{2}}, \overline{\tau}_i)$
      $D^{(-1)}.\textsc{Update}(i, \overline{\tau}_i^{-1}\nabla^2\phi_i(\overline{x}_i)^{-1}, \overline{\tau}_i)$
      $D^{(s)}.\textsc{Update}(i, \nabla^2\phi_i(\overline{x}_i)^{-\frac{1}{2}}/(\overline{\rho\tau}_i))$
  $s^{\text{tmp}}, I_s = D^{(s)}.\textsc{Query}(\rho h_s)$
  for $i \in I_s$ **do**
    if $\left\|\rho^{-1}\overline{\tau}_i^{-1}\nabla^2\phi_i(\overline{x}_i)^{-\frac{1}{2}}(s_i^{\text{tmp}} - \overline{s}_i)\right\|_2 \geq \alpha/2^{12}$ **then**
      $\overline{s}_i \leftarrow s_i^{\text{tmp}}$
      $D^{(\nabla)}.\textsc{Update}(i, \overline{s}_i/(\overline{\rho\tau}_i) + \nabla\phi_i(\overline{x}_i), \overline{x}_i, \overline{\tau}_i)$

Assumption on $\overline{\mu}$. Define

$$\widehat{\mu}_i := \frac{\overline{s}_i}{\overline{\rho\tau}_i} + \nabla\phi_i(\overline{x}_i).$$

We consider instead the difference between μ_i and $\widehat{\mu}_i$. The difference is expanded as

$$\begin{aligned}\mu_i - \widehat{\mu}_i &= \frac{s_i - \overline{s}_i}{\overline{\rho\tau}_i} + \nabla\phi_i(x_i) - \nabla\phi_i(\overline{x}_i) + \frac{s_i}{\rho\tau_i}\frac{\overline{\rho\tau}_i - \rho\tau_i}{\overline{\rho\tau}_i} \\ &= \frac{s_i - \overline{s}_i}{\overline{\rho\tau}_i} + \left(\nabla\phi_i(x_i) - \nabla\phi_i(\overline{x}_i)\right) + \frac{\overline{\rho\tau}_i - \rho\tau_i}{\overline{\rho\tau}_i}\mu_i - \frac{\overline{\rho\tau}_i - \rho\tau_i}{\overline{\rho\tau}_i}\nabla\phi_i(x_i).\end{aligned}$$

By repeating Young's inequality, it holds that

$$\begin{aligned}\|\mu_i - \widehat{\mu}_i\|^2_{\nabla^2\phi(\overline{x}_i)^{-1}} &\le 4\left\|\frac{s_i - \overline{s}_i}{\overline{\rho\tau}_i}\right\|^2_{\nabla^2\phi(\overline{x}_i)^{-1}} + 4\left\|\nabla\phi_i(x_i) - \nabla\phi_i(\overline{x}_i)\right\|^2_{\nabla^2\phi(\overline{x}_i)^{-1}} \\ &\quad + 4\left|\frac{\overline{\rho\tau}_i - \rho\tau_i}{\overline{\rho\tau}_i}\right|^2\left(\|\mu_i\|^2_{\nabla^2\phi(\overline{x}_i)^{-1}} + \|\nabla\phi_i(x_i)\|^2_{\nabla^2\phi(\overline{x}_i)^{-1}}\right) \\ &\le \alpha^2/2^{16} + \alpha^2\epsilon^2/2^8 \le \alpha^2/2^{14}.\end{aligned}$$

where we use the induction hypothesis in (29), self-concordant ϕ_i, Lemma 3, $\overline{\rho} \approx_{\alpha/2^{12}} \rho$, and also $\|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}} \le 2\Gamma_i \le 2\epsilon^2$ due to Lemma 8. Lemma 41 and the accuracy setting indicate that $\|\overline{\mu}_i - \widehat{\mu}_i\|_{\nabla^2\phi_i(\overline{x}_i)^{-1}} \le \alpha/2^7$. Consequently, the following holds by the triangle inequality

$$\|\mu_i - \overline{\mu}_i\|_{\nabla^2\phi(x_i)^{-1}} \le \|\mu_i - \widehat{\mu}_i\|_{\nabla^2\phi(x_i)^{-1}} + \|\widehat{\mu}_i - \overline{\mu}_i\|_{\nabla^2\phi(x_i)^{-1}} \le \alpha/2^5.$$

The remaining quantities in the short step have a direct correspondence. By the definitions of h_2, h_{root}, and h_r, together with the guarantees of $D^{(-1)}$ and $D^{(\text{sample})}$, the updates to the primal and slack variables in Algorithm 5 are precisely the corresponding updates prescribed by Algorithm 1,

Induction on $(\overline{x}, \overline{s}, \overline{\tau})$. It should be noticed that x is the ideal primal variable in Algorithm 1. Consequently, whenever $\overline{x}_i$ is updated, the lazy update ensures $\left\|\nabla^2\phi(x_i)^{\frac{1}{2}}(x_i^{(\text{new})} - \overline{x}_i^{(\text{new})})\right\|_2 \le \alpha/2^{18}$. For $\overline{s}$, Lemma 40 and Line 30 in Algorithm 5 ensures $\frac{1}{\overline{\rho}^{(\text{new})}\overline{\tau}_i^{(\text{new})}}\left\|\nabla^2\phi_i(\overline{x}_i^{(\text{new})})^{-\frac{1}{2}}(s_i^{\text{tmp}} - s_i^{(\text{new})})\right\|_2 \le \alpha/2^{12}$. As a consequence of the lazy update it holds that

$$\begin{aligned}&\frac{1}{\overline{\rho}^{(\text{new})}\overline{\tau}_i^{(\text{new})}}\left\|\nabla^2\phi_i(x_i^{(\text{new})})^{-\frac{1}{2}}(\overline{s}_i^{(\text{new})} - s_i^{(\text{new})})\right\|_2 \\ \le&\frac{2}{\overline{\rho\tau}_i}\left(\left\|\nabla^2\phi_i(x_i^{(\text{new})})^{-\frac{1}{2}}(s_i^{\text{tmp}} - \overline{s}_i^{(\text{new})})\right\|_2 + \left\|\nabla^2\phi_i(x_i^{(\text{new})})^{-\frac{1}{2}}(s_i^{\text{tmp}} - s_i^{(\text{new})})\right\|_2\right) \le \alpha/2^{10}.\end{aligned}$$

Moreover, Lemma 46 and Lemma 42 ensures that $\tau^{\text{tmp}} \approx_{\alpha/2^{12}} \tau(\overline{x}^{(\text{new})})$, and Corollary 25 ensures $\tau(x^{(\text{new})}) \approx_{2^8\alpha/2^{20}} \tau(\overline{x}^{(\text{new})})$. Combining with the lazy update, it holds that $\overline{\tau} \approx_{\alpha/2^{10}} \tau(x^{(\text{new})})$.

Other conditions. By Lemma 61 and the induction, our maintained Lewis weights $\overline{\tau}$ provides valid leverage-score overestimates for the inverse maintenance associated with the smoothed Lewis scaling. Moreover, Lemma 21 and Point 1 of Lemma 23 ensure the same guarantee for the Newton system $\mathbf{A}^\top\overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-1}\mathbf{A}$. To provide the solver required by Lemma 46, we follow [VDBLL+21]

and use inverse maintenance. Unlike their setting, where the existing inverse maintenance can be reused, we maintain an additional instance of $D^{(-1)}$ for the smoothing-based Gram matrices associated with $\overline{\mathbf{S}}$, where $\overline{\mathbf{S}}$ is maintained by $D^{(\text{smooth})}$. Overestimates are ensured with the same arguments and also Lemma 23.

The other conditions of Lemma 46, Lemma 47, and Lemma 48 trivially hold from the Algorithm 5. □

6.4 Path Following Analysis

Now we present the runtime analysis and the final theorem.

Theorem 54 (Restated Main Theorem)**.** *Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, vectors $b \in \mathbb{R}^d$, $c \in \mathbb{R}^n$, and convex compact sets $K_1 \ldots, K_m$ with $K_i \subseteq \mathbb{R}^{n_i}$ and $\sum_{i=1}^m n_i = n$. Suppose that $n_i = O(1)$ for any $i \in [m]$ and each K_i is equipped with a 1-self-concordant barrier, with constant-time oracle access to the value, gradient, and Hessian of ϕ_i, and to the value and gradient of its Fenchel conjugate. For any $\delta > 0$, there exists an algorithm that outputs x such that $x \in K := \prod_{i=1}^m K_i$, $\mathbf{A}^\top x = b$, and*

$$c^\top x \le \min_{\substack{x\in K\\ \mathbf{A}^\top x = b}} c^\top x + \delta$$

in time $\widetilde{O}(nd + d^{2.5})$, where $\widetilde{O}(\cdot)$ hides logarithmic terms and the polynomial of block size.

Proof. We use Point 1 in Lemma 49 to obtain an ϵ-centered initial triplet for the extended program in (28) with some δ' satisfying $\log(1/\delta') = \widetilde{O}(\log(1/\delta))$. Then we invoke BlockPathFollowing in Algorithm 4. Lemma 20, Lemma 37, and the Lemma 38 establish the convergence property, approximate feasibility, and the sampling guarantees of the corresponding ideal ShortStep in Algorithm 1. Moreover, Lemma 53 ensures that Algorithm 5 correctly implements the ideal algorithm. As a consequence, it outputs an ϵ'-centered triplet as the output of the extended program after $T = \widetilde{O}(\sqrt{d})$ calls to ShortStepImplementation. We feed the output to Lemma 50 and obtain a primal feasible solution for the extended program. Point 2 of Lemma 49 then produces some $x \in K$ such that x is approximate feasible for the original program. Also, an adaption of Lemma D.2 in [LPT+25] and Lemma 4.46 in [VDBLL+21] ensures the conditional number of $\mathbf{A}^\top\mathbf{T}^{-1}\nabla\Phi(x)^{-1}\mathbf{A}$ is bounded by $\text{poly}(n)$, allowing us to transfer from $\ell_i nfty$-norm to weighted norm of approximate feasibility. When δ' is sufficiently small, all the additional polynomial factors are absorbed, and the triplet is ϵ-central for the original program. Invoking Lemma 50 again produces a final output $x^{(\text{fi})}$ such that $c^\top x^{(\text{fi})} \le \min_{x\in K, \mathbf{A}^\top x = b} c^\top x + \delta$.

Block total updates. It requires to bound the total number of block updates for variables $\overline{x}$, $\overline{s}$, $\overline{\tau}$ in the actual implementation. We first compute for $\overline{x}$. If $\overline{x}_i$ is updated, then $\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}(x_i^{(t)} - \overline{x}_i^{(t)})\right\|_2 \ge \alpha/2^{20}$. Fix some $\alpha_{\text{nb}} \le \alpha/2^{36}$ and choose the constant C_{Var} according to the condition of Lemma 51. The same lemma indicates that there exists a nearby sequence $\widehat{x}^{(t)}$ that is α_{nb}-approximate to the ideal x_t and satisfies $\left\|\nabla^2\phi(x_i)^{\frac{1}{2}}(\widehat{x}^{(t)} - x^{(t)})\right\|_2 \le \alpha/2^{36}$. This suggests the hessian weighted norm of the difference between any two consecutive iterations when $\overline{x}_i$ is updated has lower bound $\widetilde{\Omega}(1)$. Then the total number of x-updates is bounded as

$$\widetilde{O}\left(T\sum_{t=1}^T \left\|\nabla^2\Phi(\widehat{x}^{(t)})^{\frac{1}{2}}(\widehat{x}^{(t)} - \widehat{x}^{(t-1)})\right\|_2^2\right) \le \widetilde{O}\left(\frac{Tn}{d}\sum_{t=1}^T \left\|\nabla^2\Phi(\widehat{x}^{(t)})^{\frac{1}{2}}(\widehat{x}^{(t)} - \widehat{x}^{(t-1)})\right\|_{2,\tau(\widehat{x}^{(t)})}^2\right) = \widetilde{O}(n)$$

where we use $T = \widetilde{O}(\sqrt{d})$, $\|\cdot\|_2^2 \leq \frac{m}{d}\|\cdot\|_\tau^2$, $\|\cdot\|_2 \leq \|\cdot\|_1$, and Lemma 51. This suggests $N_x = \widetilde{O}(n)$. Similar arguments give $N_s = \widetilde{O}(n)$ and $N_\tau = \widetilde{O}(n)$ as a consequence of Corollary 27, and Lemma 52.

Cost of $D^{(s)}$, $D^{(\nabla)}$, $D^{(\textbf{sample})}$, $D^{(\textbf{smooth})}$. We first consider their block update costs. Each update takes $\widetilde{O}(d)$ time according to their corresponding lemma and the precision setting in Algorithm 4. Since the total number of such operations is $\widetilde{O}(n)$, the total cost can be bounded by $\widetilde{O}(nd)$.

We now analyze the total cost of $D^{(s)}$.Query. Using Corollary 27, the amortized total cost is bounded by

$$\widetilde{O}\left(Td + dT\sum_{t=1}^{T}\left\|\mathbf{W}^{(t)}(s^{(t)} - s^{(t-1)})\right\|_2^2\right) \leq \widetilde{O}\left(Td + mT^2\right) \leq \widetilde{O}(nd),$$

where $\mathbf{W}^{(t)}$ is $\overline{\rho}^{-1}\overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}$ at t-th iteration. Also, the total cost of $D^{(\nabla)}$.Query is simply $\widetilde{O}(nT) \leq \widetilde{O}(n\sqrt{d})$. Finally, for total cost of $D^{(\text{sample})}$.Sample, Lemma 38 indicates that each call to Sample needs $\widetilde{O}(m/\sqrt{d} + d)$ independent copies, and each copy requires $\widetilde{O}(d)$ cost. Then the total cost across $\widetilde{O}(\sqrt{d})$ iteration is then $\widetilde{O}(nd + d^{2.5})$.

Cost of inverse maintenance. The necessary conditions for Lemma 48 hold due to Lemma 51, Lemma 52, and the implementation. The above computation suggests there are $\widetilde{O}(n)$ calls to $D^{(-1)}$.Update, then the total cost of block updates is bounded by $\widetilde{O}(n)$. By Lemma 48, the total cost of $D^{(-1)}$.Solve is

$$\widetilde{O}\left(nd + d^\omega + T\left(n + d^2 + d^{\omega - 1/2}\right)\right) = \widetilde{O}(nd + d^{5/2})$$

as the matrix multiplication exponent has $\omega < 5/2$.

Cost of $D^{(\tau)}$. Lemma 52 ensures the necessary condition of Lemma 46. Similar to [VDBLL+21], the additional inverse maintenance condition required by Lemma 46 can be hold as a consequence of similar arguments and Lemma 23. We consider the total cost of $D^{(\tau)}$.ScaleBlock(). Fix i, each update of i-th block takes $\widetilde{O}(n\tau_i)$ time by Lemma 46 and is updated only when $\overline{x}_i$ changes. Similar to the previous N_x analysis, the total block update time is

$$\widetilde{O}\left(nT\sum_{t=1}^{T}\left\|\nabla\Phi(\widehat{x}^{(t)})^{\frac{1}{2}}(\widehat{x}^{(t)} - \widehat{x}^{(t-1)})\right\|_{\tau(\widehat{x}^{(t)})}^2\right) = \widetilde{O}(nd)$$

Moreover, Lemma 46 bounds the total cost of $D^{(\tau)}$.Query by

$$\widetilde{O}\left(Td^2 + T^2m + TT_{\text{inverse}}\right) = \widetilde{O}(nd + d^{5/2}).$$

where we use the cost of D^{-1}.Query as in Lemma 48.

Cost of direct x computation. It remains to bound the cost of x formulation in Algorithm 5. We notice that it requires $\widetilde{O}(d)$ time for every sample block. The block-diagonal sampling scheme in Lemma 38 ensures that at most $\widetilde{O}(m/\sqrt{d}+d)$ blocks are sampled each iteration. This gives total $\widetilde{O}(nd+d^{2.5})$ total time. The total time of dense gradient computation and other direct maintenance is also covered by $\widetilde{O}(nd)$. Fast linear system solvers in [NN13, CLM+15] give $\widetilde{O}(nd+d^{\omega})$ time for the initial and final processing. Therefore the total running time is

$$\widetilde{O}(nd+d^{5/2})$$

where $\widetilde{O}$ hides all polylgarithmic factors. □

AI Use Disclosure. Generative AI tools were used for language editing, improving the presentation of proofs, and proof checking. The IPM analysis was developed by the authors, with AI assistance limited to writing refinement and proof auditing. For the data-structure implementation, AI tools also provided limited assistance in extending certain existing constructions to the block setting. All AI-assisted mathematical arguments and algorithmic extensions were independently verified by the authors, who take full responsibility for the content and correctness of the paper.

References

[AGS24] Simon Apers, Sander Gribling, and Aaron Sidford. On computing approximate lewis weights, 2024. arXiv preprint.

[AJK25] Deeksha Adil, Shunhua Jiang, and Rasmus Kyng. Acceleration meets inverse maintenance: Faster ℓ_∞-regression. In *Proc. 52nd International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 5:1–5:16, 2025.

[AMV21] Kyriakos Axiotis, Aleksander Madry, and Adrian Vladu. Faster sparse minimum cost flow by electrical flow localization. In *Proc. 62nd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 528–539, 2021.

[AvdBGZ24] Emile Anand, Jan van den Brand, Mehrdad Ghadiri, and Daniel J. Zhang. The bit complexity of dynamic algebraic formulas and their determinants. In *Proc. 51st International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 10:1–10:20, 2024.

[CJN18] Michael B. Cohen, T. S. Jayram, and Jelani Nelson. Simple analyses of the sparse Johnson–Lindenstrauss transform. In *Proc. 1st Symposium on Simplicity in Algorithms (SOSA)*, pages 15:1–15:9, 2018.

[CKL+22] Li Chen, Rasmus Kyng, Yang P. Liu, Richard Peng, Maximilian Probst Gutenberg, and Sushant Sachdeva. Maximum flow and minimum-cost flow in almost-linear time. In *Proc. 63rd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 612–623, 2022.

[CLM+15] Michael B Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In *Proceedings of the 2015 conference on innovations in theoretical computer science*, pages 181–190, 2015.

[CLS19] Michael B. Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. In *Proc. 51st ACM Symposium on Theory of Computing (STOC)*, pages 938–942, 2019.

[CMSV17] Michael B. Cohen, Aleksander Madry, Piotr Sankowski, and Adrian Vladu. Negative-weight shortest paths and unit capacity minimum cost flow in $\widetilde{O}(m^{10/7} \log w)$ time. In *Proc. 28th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 752–771, 2017.

[CP15] Michael B. Cohen and Richard Peng. ℓ_p row sampling by lewis weights. In *Proc. 47th ACM Symposium on Theory of Computing (STOC)*, pages 183–192, 2015.

[DBW12] John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

[FLPS22] Maryam Fazel, Yin Tat Lee, Swati Padmanabhan, and Aaron Sidford. Computing lewis weights to high precision. In *Proc. 33rd ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2723–2742, 2022.

[GPV23] Mehrdad Ghadiri, Richard Peng, and Santosh S. Vempala. The bit complexity of efficient continuous optimization. In *Proc. 64th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 2059–2070, 2023.

[GSZ26] Sander Gribling, Aaron Sidford, and Chenyi Zhang. Computing lewis weights to high precision using local relative smoothness. In *Proc. 39th Conference on Learning Theory (COLT)*, pages 2901–2939, 2026.

[HJS+22] Baihe Huang, Shunhua Jiang, Zhao Song, Runzhou Tao, and Ruizhe Zhang. Solving SDP faster: A robust IPM framework and efficient implementation. In *Proc. 63rd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 233–244, 2022.

[JKL+20] Haotian Jiang, Tarun Kathuria, Yin Tat Lee, Swati Padmanabhan, and Zhao Song. A faster interior point method for semidefinite programming. In *Proc. 61st IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 910–918, 2020.

[JKLS25] Shunhua Jiang, Michael Kapralov, Lawrence Li, and Aaron Sidford. Generalized flow in nearly-linear time on moderately dense graphs. In *Proc. 66th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1091–1099, 2025.

[JL84] William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. In *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.

[JLLS23] Arun Jambulapati, James R. Lee, Yang P. Liu, and Aaron Sidford. Sparsifying sums of norms. In *Proc. 64th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1953–1962, 2023.

[JLLS24] Arun Jambulapati, James R. Lee, Yang P. Liu, and Aaron Sidford. Sparsifying generalized linear models. In *Proc. 56th ACM Symposium on Theory of Computing (STOC)*, pages 1665–1675, 2024.

[JNW22] Shunhua Jiang, Bento Natura, and Omri Weinstein. A faster interior-point method for sum-of-squares optimization. In *Proc. 49th International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 79:1–79:20, 2022.

[JPW23] Shunhua Jiang, Binghui Peng, and Omri Weinstein. The complexity of dynamic least-squares regression. In *Proc. 64th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1605–1627, 2023.

[JSWZ21] Shunhua Jiang, Zhao Song, Omri Weinstein, and Hengjie Zhang. A faster algorithm for solving general LPs. In *Proc. 53rd ACM Symposium on Theory of Computing (STOC)*, pages 823–832, 2021.

[Kar84] Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311, 1984.

[KN14] Daniel M. Kane and Jelani Nelson. Sparser Johnson–Lindenstrauss transforms. *Journal of the ACM*, 61(1), 2014.

[KNPW11] Daniel M. Kane, Jelani Nelson, Ely Porat, and David P. Woodruff. Fast moment estimation in data streams in optimal space. In *Proc. 43rd ACM Symposium on Theory of Computing (STOC)*, pages 745–754, 2011.

[LNNT16] Kasper Green Larsen, Jelani Nelson, Huy L. Nguyen, and Mikkel Thorup. Heavy hitters via cluster-preserving clustering. In *Proc. 57th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 61–70, 2016.

[LPT+25] Yang P Liu, Richard Peng, Colin Tang, Albert Weng, and Junzhao Yang. Adaptive matrix sparsification and applications to empirical risk minimization. *arXiv preprint arXiv:2512.02003*, 2025.

[LS14] Yin Tat Lee and Aaron Sidford. Path finding methods for linear programming: Solving linear programs in o (vrank) iterations and faster algorithms for maximum flow. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 424–433. IEEE, 2014.

[LS15] Yin Tat Lee and Aaron Sidford. Efficient inverse maintenance and faster algorithms for linear programming. In *Proc. 56th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 230–249, 2015.

[LS19] Yin Tat Lee and Aaron Sidford. Solving linear programs with sqrt (rank) linear system solves. *arXiv preprint arXiv:1910.08033*, 2019.

[LS20] Yang P. Liu and Aaron Sidford. Faster energy maximization for faster maximum flow. In *Proc. 52nd ACM Symposium on Theory of Computing (STOC)*, pages 803–814, 2020.

[LSZ19] Yin Tat Lee, Zhao Song, and Qiuyi Zhang. Solving empirical risk minimization in the current matrix multiplication time. In *Conference on Learning Theory*, pages 2140–2157. PMLR, 2019.

[Mad13] Aleksander Madry. Navigating central path with electrical flows: From flows to matchings, and back. In *Proc. 54th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 253–262, 2013.

[MO25] Naren Sarayu Manoj and Max Ovsiankin. The change-of-measure method, block lewis weights, and approximating matrix block norms. In *Proc. 36th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2252–2300, 2025.

[MP26] Naren Sarayu Manoj and Kumar Kshitij Patel. Distributionally robust linear regression with block lewis weights. In *Proc. International Conference on Learning Representations (ICLR)*, 2026.

[Nes13] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013.

[NN13] Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 ieee 54th annual symposium on foundations of computer science*, pages 117–126. IEEE, 2013.

[NS19] Vasileios Nakos and Zhao Song. Stronger ℓ_2/ℓ_2 compressed sensing; without iterating. In *Proc. 51st ACM Symposium on Theory of Computing (STOC)*, pages 289–297, 2019.

[NSW19] Vasileios Nakos, Zhao Song, and Zhengyu Wang. (nearly) sample-optimal sparse fourier transform in any dimension; RIPless and filterless. In *Proc. 60th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1568–1577, 2019.

[Pad26] Swati Padmanabhan. Computing lewis weights to high precision by fixed-point iteration, 2026. arXiv preprint.

[Pag13] Rasmus Pagh. Compressed matrix multiplication. *ACM Transactions on Computation Theory*, 5(3):9:1–9:17, 2013.

[PSW17] Eric Price, Zhao Song, and David P. Woodruff. Fast regression with an ℓ_∞ guarantee. In *Proc. 44th International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 59:1–59:14, 2017.

[Ren88] James Renegar. A polynomial-time algorithm, based on newton's method, for linear programming. *Mathematical programming*, 40(1):59–93, 1988.

[San04] Piotr Sankowski. Dynamic transitive closure via dynamic matrix inverse (extended abstract). In *Proc. 45th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 509–517, 2004.

[SY20] Zhao Song and Zheng Yu. Oblivious sketching-based central path method for solving linear programming problems, 2020.

[Tro15] Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and trends® in machine learning*, 8(1-2):1–230, 2015.

[Vai89] Pravin M Vaidya. A new algorithm for minimizing convex functions over convex sets. In *30th Annual Symposium on Foundations of Computer Science*, pages 338–343. IEEE Computer Society, 1989.

[vdB20] Jan van den Brand. A deterministic linear program solver in current matrix multiplication time. In *Proc. 31st ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 259–278, 2020.

[vdB21] Jan van den Brand. Unifying matrix data structures: Simplifying and speeding up iterative algorithms. In *Proc. Symposium on Simplicity in Algorithms (SOSA)*, pages 1–13, 2021.

[vdBCP+23] Jan van den Brand, Li Chen, Richard Peng, Rasmus Kyng, Yang P. Liu, Maximilian Probst Gutenberg, Sushant Sachdeva, and Aaron Sidford. A deterministic almost-linear time algorithm for minimum-cost flow. In *Proc. 64th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 503–514, 2023.

[VDBLL+21] Jan Van Den Brand, Yin Tat Lee, Yang P Liu, Thatchaphol Saranurak, Aaron Sidford, Zhao Song, and Di Wang. Minimum cost flows, mdps, and l1-regression in nearly linear time for dense instances. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 859–869, 2021.

[vdBLN+20] Jan van den Brand, Yin-Tat Lee, Danupon Nanongkai, Richard Peng, Thatchaphol Saranurak, Aaron Sidford, Zhao Song, and Di Wang. Bipartite matching in nearly-linear time on moderately dense graphs. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 919–930. IEEE, 2020.

[vdBLSS20] Jan van den Brand, Yin Tat Lee, Aaron Sidford, and Zhao Song. Solving tall dense linear programs in nearly linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 775–788, 2020.

[vdBNS19] Jan van den Brand, Danupon Nanongkai, and Thatchaphol Saranurak. Dynamic matrix inverse: Improved algorithms and matching conditional lower bounds. In *Proc. 60th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 456–480, 2019.

A Basic Mathematical Facts and Proofs

A.1 Proofs involving Self-concordance, Fenchel Conjugate and Bregman Divergence

We provide the proofs for the basic facts on self-concordance, Fenchel Conjugate, and Bregman Divergence presented in Section 3.

Proof of Lemma 3. The first two points are standard facts of self-concordant barriers [Nes13]. The last two follow by integrating the Hessian along the segment $x + th$ for $t \in [0, 1]$ and applying the Hessian comparison in Point 2. □

Proof of Lemma 5. By the 1st-order optimality of the Fenchel conjugate, we have

$$y = \nabla\phi(x(y)), \qquad x(y) = (\nabla\phi)^{-1}(y).$$

We take gradient for both sides of $\phi^*(y) = \langle x(y), y\rangle - \phi(x(y))$ and get $\nabla\phi^*(y) = x(y)$. In a similar fashion, taking gradients of $y = \nabla\phi(x(y))$ gives $\nabla^2\phi^*(y) = \frac{\mathrm{d}x(y)}{\mathrm{d}y} = \nabla^2\phi(x(y))^{-1}$. □

Proof of Lemma 6. It suffices to show that ϕ^* is also a self-concordant function and then we can invoke Lemma 3 to obtain the bounds. Let $x = \nabla\phi^*(a)$ and $\mathbf{H} = \nabla^2\phi(x)$. By Lemma 5, we have $\nabla^2\phi^*(a) = \mathbf{H}^{-1}$. As a consequence, we obtain

$$\left|\nabla^3\phi^*(a)[h,h,h]\right| = \left|\nabla^3\phi(x)[\mathbf{H}^{-1}h, \mathbf{H}^{-1}h, \mathbf{H}^{-1}h]\right| \le 2\left\|h\right\|^3_{\nabla^2\phi^*(a)}$$

Therefore ϕ^* is also self-concordant. □

Proof of Lemma 8. First, the Bregman divergence $D_{\phi^*}(a,b)$ is expanded in a second-order integral representation

$$D_{\phi^*}(a,b) = \phi^*(a) - \phi^*(b) - \langle\nabla\phi^*(b), a-b\rangle = \int_0^1 (1-t)\langle\nabla^2\phi^*(b+t(a-b))(a-b),(a-b)\rangle\mathrm{d}t.$$

The local norm r admits a similar integral form

$$\begin{aligned} r^2 = \|a-b\|^2_{\nabla^2\phi^*(b)} &= \|a\|^2_{\nabla^2\phi^*(b)} - \|b\|^2_{\nabla^2\phi^*(b)} - 2\left\langle\nabla^2\phi^*(b)b, a-b\right\rangle \\ &= 2\int_0^1 (1-t)\langle\nabla^2\phi^*(b)(a-b),(a-b)\rangle\mathrm{d}t. \end{aligned}$$

As a result, we compute

$$\begin{aligned} \left|2D_{\phi^*}(a,b) - r^2\right| &= 2\left|\int_0^1 (1-t)\left\langle\nabla^2\phi^*(b+t(a-b)) - \nabla^2\phi^*(b), (a-b)(a-b)^\top\right\rangle\mathrm{d}t\right| \\ &\le 2\int_0^1 (1-t)\left\|\nabla^2\phi^*(b)^{-\frac{1}{2}}\nabla^2\phi^*(b+t(a-b))\nabla^2\phi^*(b)^{-\frac{1}{2}} - \mathbf{I}\right\|_{\mathrm{op}}\|a-b\|^2_{\nabla^2\phi^*(b)}\,\mathrm{d}t \\ &\le 2\int_0^1 \frac{(1-t)3t\left\|a-b\right\|^3_{\nabla^2\phi^*(b)}}{1-t\left\|a-b\right\|_{\nabla^2\phi^*(b)}}\mathrm{d}t \\ &\le 2\left\|a-b\right\|^3_{\nabla^2\phi^*(b)} \end{aligned}$$

where we use self-concordance, Lemma 3, and Lemma 5. □

A.2 Projection Matrix facts

We provide some analysis tools for projection matrix, and the block leverage scores.

Lemma 55. *Let $\mathbf{P} \in \mathbb{R}^{n\times n}$ be any projection matrix. Let $\mathbf{\Sigma} = \mathrm{tr}_{\mathrm{G}}(\mathbf{P}) \in \mathbb{R}^{m\times m}$.*

(i) $\mathbf{\Sigma}_{ii} = \sum_{j\in[m]} (\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})_{ij}$;

(ii) $\mathbf{0} \preceq \mathbf{P} \circ_{\mathrm{G}} \mathbf{P} \preceq \mathbf{\Sigma}$;

(iii) $(\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})\mathbf{1} = \mathbf{\Sigma}\mathbf{1}$;

(iv) $\left\|\mathbf{\Sigma}^{-1}(\mathbf{P} \circ_G \mathbf{P})x\right\|_\infty \le \|x\|_\infty$

Proof. For (i), we use the definition of $\circ_{\mathrm{G}}$ and projection matrix $\mathbf{P} = \mathbf{PP}$ to obtain

$$\mathbf{\Sigma}_{ii} = \mathrm{tr}(\mathbf{P}_{ii}) = \mathrm{tr}((\mathbf{PP})_{ii}) = \mathrm{tr}\left(\sum_{j\in[m]} \mathbf{P}_{ij}\mathbf{P}_{ji}\right) = \sum_{j\in[m]} \mathrm{tr}(\mathbf{P}_{ij}\mathbf{P}_{ji}) = \sum_{j\in[m]} (\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})_{ij}.$$

For (ii), we first notice, since $\mathbf{P}$ is a projection matrix, all of its eigenvalues are either 0 or 1. Hence, we have $\mathbf{I} - \mathbf{P} \succeq \mathbf{0}$ and $\mathbf{P} \succeq 0$. We invoke a block version of Schurr product theorem to obtain $\mathbf{\Sigma} - \mathbf{P} \circ_{\mathrm{G}} \mathbf{P} = \mathbf{P} \circ_{\mathrm{G}} (\mathbf{I} - \mathbf{P}) \succeq 0$.

To prove (iv), it suffices to show for any $x \in \mathbb{R}^m$,

$$\left|e_i^\top (\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})x\right| = \left|\sum_j (\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})_{ij} x_j\right| \le \sum_j (\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})_{ij} |x_j| = \mathbf{\Sigma}_{ii} |x_j| \le \mathbf{\Sigma}_{ii} \|x\|_\infty$$

where the second step is due to $(\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})_{ij} = \|\mathbf{P}_{ij}\|_{\mathrm{F}}^2 \ge 0$ and the third is an application of (i). □

Lemma 56. *Let $p \in (0, 2]$. For any two symmetric and block-diagonal SPD matrices $\mathbf{C}, \mathbf{C}' \in \mathbb{R}^{n\times n}$, if $(\mathbf{C}'_i)^2 \approx_\epsilon \mathbf{C}_i^2$ is true for some $\epsilon > 0$ and any block $i \in [m]$, then for any nonnegative v,*

$$\|v\|_{\mathbf{Q}} \le e^{2\epsilon} \|v\|_{\mathbf{Q}'}.$$

where $\mathbf{Q} := \mathbf{P}(\mathbf{CA}) \circ_{\mathrm{G}} \mathbf{P}(\mathbf{CA})$, and $\mathbf{Q}' := \mathbf{P}(\mathbf{C}'\mathbf{A}) \circ_{\mathrm{G}} \mathbf{P}(\mathbf{C}'\mathbf{A})$.

Proof. We define $\mathbf{V} = \mathsf{diag}_{\mathrm{G}}(v)$, $\mathbf{M} = \mathbf{A}^\top\mathbf{C}^2\mathbf{A}$, and $\mathbf{N} = \mathbf{A}^\top\mathbf{CVCA}$. Similarly, $\mathbf{M}', \mathbf{N}'$ are defined analogously. Since $\mathbf{V}$ is a block-identity matrix and nonnegative, it holds that

$$\mathbf{M}' \approx_\epsilon \mathbf{M}, \qquad \mathbf{N}' \approx_\epsilon \mathbf{N}.$$

By the definition of the block Schur product,

$$\|v\|_{\mathbf{Q}}^2 = \mathrm{tr}\left(\mathbf{VP}(\mathbf{CA})\mathbf{VP}(\mathbf{CA})\right) = \mathrm{tr}(\mathbf{M}^{-1}\mathbf{NM}^{-1}\mathbf{N}).$$

It holds that $\mathrm{tr}(\mathbf{M}^{-1}\mathbf{NM}^{-1}(e^\epsilon\mathbf{N}' - \mathbf{N})) \ge 0$, as the trace of the product of two PSD matrices is nonnegative. By the cycle property of trace and switching the position of $\mathbf{M}$ and $\mathbf{N}$, we obtain

$$\begin{aligned}\|v\|_{\mathbf{Q}}^2 = \mathrm{tr}(\mathbf{M}^{-1}\mathbf{NM}^{-1}\mathbf{N}) &\le e^{2\epsilon}\mathrm{tr}(\mathbf{M}^{-1}\mathbf{N}'\mathbf{M}^{-1}\mathbf{N}') \\ &\le e^{4\epsilon}\mathrm{tr}((\mathbf{M}')^{-1}\mathbf{N}'(\mathbf{M}')^{-1}\mathbf{N}') = e^{4\epsilon}\|v\|_{\mathbf{Q}'}^2.\end{aligned}$$

The result follows by taking square root. □

A.3 Block Lewis-weight Facts

We provide the differential properties of the block Lewis-weight. The analysis and notations are direct extension of the well-established analysis in [LS14] and [LS19]. This first lemma is the variational formulation of Lewis weights. This is almost trivial from matrix calculus and we omit the proof.

Lemma 57. *For $p \in (0,2)$, the v-regularized block Lewis weight of p-th order satisfies the following:*

$$w(\mathbf{C}) := \mathrm{argmin}_{w \in \mathbb{R}^m_{>0}} f(w, \mathbf{C})$$

where

$$f(w,\mathbf{C}) = -\frac{1}{1-2/p}\log\det(\mathbf{A}^\top \mathbf{C}\mathbf{W}^{1-2/p}\mathbf{C}\mathbf{A}) + \sum_{i=1}^m w_i - \sum_{i=1}^m v_i \log w_i$$

with $\mathbf{W} = \mathsf{diag}_{\mathrm{G}}(w)$.

We compute the first and second-order derivatives of $f(w,\mathbf{C})$. It requires the following fact from [LS19].

Lemma 58 (Matrix form of Lemma 49 in [LS19])**.** *Given a full rank $\mathbf{A} \in \mathbb{R}^{n\times d}$, symmetric $\mathbf{B}, \Delta \in \mathbb{R}^{n\times n}$, it holds that*

$$\mathrm{D}_{\mathbf{B}}\mathbf{P}(\mathbf{B}\mathbf{A})[\Delta] = \Delta\mathbf{B}^{-1}\mathbf{P}(\mathbf{B}\mathbf{A}) + \mathbf{P}(\mathbf{B}\mathbf{A})\mathbf{B}^{-1}\Delta - \mathbf{P}(\mathbf{B}\mathbf{A})\left(\Delta\mathbf{B}^{-1} + \mathbf{B}^{-1}\Delta\right)\mathbf{P}(\mathbf{B}\mathbf{A}).$$

Lemma 59. *The first and second-order derivative of $f(w,\mathbf{C})$ has the following form*

$$\nabla_w f(w,\mathbf{C}) = \mathbf{1}_m - \mathsf{diag}(w)^{-1}\left(\mathrm{tr}_{\mathrm{G}}\left(\mathbf{P}_C\right) + v\right),$$

and

$$\nabla_{ww} f(w,\mathbf{C}) = \mathsf{diag}(w)^{-1}\left(\mathbf{\Sigma}_C + \mathbf{V} - \left(1-\frac{2}{p}\right)\cdot \mathbf{P}_C \circ_{\mathrm{G}} \left(\mathbf{I}_n - \mathbf{P}_C\right)\right)\mathsf{diag}(w)^{-1}.$$

where $\mathbf{P}_C = \mathbf{P}(\mathbf{W}^{1/2-1/p}\mathbf{C}\mathbf{A})$, $\mathbf{\Sigma}_C = \mathsf{diag}(\mathrm{tr}_{\mathrm{G}}(\mathbf{P}_C))$.

Proof. Let $\mathbf{B} = \mathbf{W}^{1/2-1/p}\mathbf{C}$ with fixed $\mathbf{C}$. It then holds that $\mathrm{d}\mathbf{B}\mathbf{B}^{-1} = \left(\frac{1}{2}-\frac{1}{p}\right)\mathsf{diag}_{\mathrm{G}}(w^{-1}\mathrm{d}w)$. We insert Lemma 58 to compute

$$\mathrm{d}\mathbf{P} = \left(1-\frac{2}{p}\right)\left(\frac{1}{2}\mathsf{diag}_{\mathrm{G}}(w^{-1}\mathrm{d}w)\mathbf{P} + \frac{1}{2}\mathbf{P}\mathsf{diag}_{\mathrm{G}}(w^{-1}\mathrm{d}w) - \mathbf{P}\mathsf{diag}_{\mathrm{G}}(w^{-1}\mathrm{d}w)\mathbf{P}\right).$$

As a consequence, we have

$$\mathrm{d}\nabla_w f(w,\mathbf{C}) = \mathsf{diag}(w^{-2}\mathrm{d}w)\mathrm{tr}_{\mathrm{G}}(\mathbf{P}_C) - \mathsf{diag}(w^{-1})\mathrm{tr}_{\mathrm{G}}(\mathrm{d}\mathbf{P}_C) + \mathsf{diag}(w^{-2}\mathrm{d}w)v$$

and also

$$\begin{aligned}\nabla_{ww} f(w,\mathbf{C}) &= \mathsf{diag}(w^{-1})\left(\mathrm{tr}_{\mathrm{G}}(\mathbf{P}_C) - \left(1-\frac{2}{p}\right)\left(\mathrm{tr}_{\mathrm{G}}(\mathbf{P}_C) - \mathbf{P}_C \circ_{\mathrm{G}} \mathbf{P}_C\right)\right)\mathsf{diag}(w^{-1}) + \mathsf{diag}(w^{-2})v \\ &= \mathsf{diag}(w^{-1})\left(\mathbf{\Sigma}_C + \mathbf{V} - \left(1-\frac{2}{p}\right)\left(\mathbf{P}_C \circ_{\mathrm{G}} \left(\mathbf{I}_n - \mathbf{P}_C\right)\right)\right)\mathsf{diag}(w^{-1}).\end{aligned}$$

□

Lemma 60. *Let $\mathbf{C}, \mathbf{\Delta}_C \in \mathbb{R}^{n\times n}$ be symmetric block-diagonal matrices such that $\mathbf{C}_t = \mathbf{C} + t\mathbf{\Delta}_C$ is SPD. Then we have the following time derivative*

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla_w f(w, \mathbf{C}_t) = -2\mathsf{diag}(w^{-1})\mathrm{tr}_{\mathrm{G}}\left(\mathbf{P}(\mathbf{W}^{1/2-1/p}\mathbf{C}_t\mathbf{A})\mathrm{sym}(\mathbf{C}_t^{-1}\mathbf{\Delta}_C)(\mathbf{I}_n - \mathbf{P}(\mathbf{W}^{1/2-1/p}\mathbf{C}_t\mathbf{A}))\right).$$

Proof. We slightly abuse the notation and denote $\mathbf{P}_t = \mathbf{P}(\mathbf{W}^{1/2-1/p}\mathbf{C}_t\mathbf{A})$ when $\mathbf{W}$ is fixed. Let $\mathbf{V} = \mathbf{W}^{1/2-1/p}\mathbf{C}_t$. It holds that $\mathrm{d}\mathbf{V}\mathbf{V}^{-1} = \mathrm{d}t\mathbf{\Delta}_C\mathbf{C}_t^{-1}$. We invoke the Frechet derivative of projection matrix in Lemma 58 to obtain

$$\begin{aligned}
&\mathrm{d}\nabla_w f(w, \mathbf{C}_t) = -\mathsf{diag}(w^{-1})\mathrm{tr}_{\mathrm{G}}\left(\mathrm{d}\mathbf{P}_t\right)\\
=& -\mathsf{diag}(w^{-1})\mathrm{tr}_{\mathrm{G}}\Big(\mathrm{d}t\mathbf{\Delta}_C\mathbf{C}_t^{-1}\mathbf{P}_t + \mathbf{P}_t\mathbf{C}_t^{-1}\mathbf{\Delta}_C\mathrm{d}t - \mathbf{P}_t\left(\mathrm{d}t\mathbf{\Delta}_C\mathbf{C}_t^{-1} + \mathbf{C}_t^{-1}\mathbf{\Delta}_C\mathrm{d}t\right)\mathbf{P}_t\Big).
\end{aligned}$$

We compute the total derivative w.r.t time t:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\nabla_w f(w, \mathbf{C}_t) &= -\mathsf{diag}(w^{-1})\mathrm{tr}_{\mathrm{G}}\Big(\mathbf{\Delta}_C\mathbf{C}_t^{-1}\mathbf{P}_t + \mathbf{P}_t\mathbf{C}_t^{-1}\mathbf{\Delta}_C - \mathbf{P}_t\left(\mathbf{\Delta}_C\mathbf{C}_t^{-1} + \mathbf{C}_t^{-1}\mathbf{\Delta}_C\right)\mathbf{P}_t\Big)\\
&= -2\mathsf{diag}(w^{-1})\mathrm{tr}_{\mathrm{G}}\left(\mathbf{P}_t\mathrm{sym}(\mathbf{C}_t^{-1}\mathbf{\Delta}_C) - \mathbf{P}_t\mathrm{sym}(\mathbf{C}_t^{-1}\mathbf{\Delta}_C)\mathbf{P}_t\right)\\
&= -2\mathsf{diag}(w^{-1})\mathrm{tr}_{\mathrm{G}}\left(\mathbf{P}_t\mathrm{sym}(\mathbf{C}_t^{-1}\mathbf{\Delta}_C)(\mathbf{I}_n - \mathbf{P}_t)\right)
\end{aligned}$$

where the second equality is due to cyclic property of tr_{G}. □

Finally, we present a result to show Lewis weight is an overestimates of corresponding block leverage score.

Lemma 61. *Let $\tau = w_p(\mathbf{A})$ be the p-th block Lewis weight with $p = 1 - \theta$. If $\theta \le 1/(8\log m)$ satisfies, then the following property holds that for every $i \in [m]$*

$$\tau_i \gtrsim \sigma_i(\mathbf{T}^{-\frac{1}{2}}\mathbf{A}).$$

Proof. We reload and define a local notation $\mathbf{H}_p := \left(\mathbf{A}^\top\mathbf{T}^{1-\frac{2}{p}}\mathbf{A}\right)^{-1}$. From the definition of the block leverage score, it holds that

$$\sigma_i\left(\mathbf{T}^{\frac{1}{2}-\frac{1}{p}}\mathbf{A}\right) = \tau_i^{1-\frac{2}{p}}\mathrm{tr}\left(\mathbf{H}_p\mathbf{A}_i^\top\mathbf{A}_i\right), \qquad \sigma_i\left(\mathbf{T}^{-\frac{1}{2}}\mathbf{A}\right) = \tau_i^{-1}\mathrm{tr}\left(\mathbf{H}_1\mathbf{A}_i^\top\mathbf{A}_i\right).$$

It requires to establish the Lowner order relationship between $\mathbf{H}_p$ and $\mathbf{H}_1$.

$$\mathbf{H}_p^{-1} = \sum_i \tau_i^{1-2/p}\mathbf{A}_i^\top\mathbf{A}_i \preceq (m/d)^{\frac{2-2p}{p}}\sum_i \tau_i^{-1}\mathbf{A}_i^\top\mathbf{A}_i$$

due to $\tau_i \ge d/m$. Consequently, $\mathbf{H}_1 \preceq (m/d)^{\frac{2-2p}{p}}\mathbf{H}_p$. We insert this back and obtain

$$\begin{aligned}
\sigma_i\left(\mathbf{T}^{-\frac{1}{2}}\mathbf{A}\right) = \tau_i^{-1}\mathrm{tr}\left(\mathbf{H}_1\mathbf{A}_i^\top\mathbf{A}_i\right) &\le (m/d)^{\frac{2-2p}{p}}\tau_i^{-1}\mathrm{tr}\left(\mathbf{H}_p\mathbf{A}_i^\top\mathbf{A}_i\right)\\
&= (m\tau_i/d)^{\frac{2-2p}{p}}\sigma_i\left(\mathbf{T}^{\frac{1}{2}-\frac{1}{p}}\mathbf{A}\right) \le \tau_i(2m)^{\frac{2-2p}{p}} \le O(\tau_i)
\end{aligned}$$

where the second inequality comes from $\tau_i \le \sum_{i=1}^m \tau_i \le 2d$, and the final one is due to $\theta \le 1/(8\log m)$. □

B Proofs of IPM Analysis

B.1 Stability of x, s and $\mathbf{C}$

This subsection contains the deferred proof for the stability of x, s, $\mathbf{C}$.

Proof of Lemma 26. We present the proof for the four claims on δ_x in a sequential style. The norm bounds of g, δ_1, and δ_2 are byproducts of the proof and will simply be noted when they arise. Also, since $\overline{x}$ is an α-approximation of x, it holds that $\nabla^2\phi_i(x_i) \approx_{4\alpha} \nabla^2\phi_i(\overline{x}_i)$ as a consequence of Point 2 in Lemma 3. We will frequently use the fact in the proof without explicit referencing.

Bound on $\left\|\nabla^2\Phi(x)^{\frac{1}{2}}\mathbb{E}[\delta_x]\right\|_{2,\tau+\infty}$. We first decompose the expectation of δ_x using vectors δ_1, δ_2 and g in Algorithm 1

$$\mathbb{E}[\delta_x] = \nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbb{E}[g - \mathbf{R}\delta_r] = \nabla^2\Phi(\overline{x})^{-\frac{1}{2}}(g - \delta_1) - \nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\delta_2 \tag{30}$$

It requires to bound the blockwise ℓ_2 norms . We start with the i-th block of last term in (30). It holds that

$$\left\|(\nabla^2\Phi(x)^{\frac{1}{2}}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\delta_2)_i\right\|_2 = \left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\nabla^2\phi_i(\overline{x}_i)^{-\frac{1}{2}}\delta_{2,i}\right\| \leq e^{2\alpha}\left\|\delta_{2,i}\right\|_2.$$

In particular, we compute

$$\begin{aligned}
\left\|\delta_{2,i}\right\|_2 &= \left\|\mathbf{E}_i\overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}(\mathbf{A}^\top x - b)\right\|_2 \\
&\leq \left\|\mathbf{A}^\top x - b\right\|_{\overline{\mathbf{H}}^{-1}}\left(\lambda_{\max}\left(\mathbf{E}_i\overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-1}\mathbf{E}_i\right)\right)^{\frac{1}{2}} \\
&\leq \left\|\mathbf{A}^\top x - b\right\|_{\overline{\mathbf{H}}^{-1}}\cdot\overline{\tau}_i^{-\frac{1}{2}}\left(\mathrm{tr}\left(\mathbf{E}_i\overline{\mathbf{T}}^{-\frac{1}{2}}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-\frac{1}{2}}\mathbf{E}_i\right)\right)^{\frac{1}{2}} \\
&\leq e^{2\beta}\left\|\mathbf{A}^\top x - b\right\|_{\overline{\mathbf{H}}^{-1}}\cdot\overline{\tau}_i^{-\frac{1}{2}}\left(\mathrm{tr}\left(\mathbf{P}\left(\overline{\mathbf{T}}^{-\frac{1}{2}}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\right)_{ii}\right)\right)^{\frac{1}{2}}
\end{aligned}$$

where the first line comes from a weighted Cauchy-Schwarz, the second line from $\lambda_{\max}(\mathbf{X}^\top\mathbf{Y}\mathbf{X}) \leq \left\|\mathbf{X}^\top\mathbf{X}\right\|_2\lambda_{\max}(\mathbf{Y})$ with $\mathbf{Y} \succeq 0$, and the rests are due to spectral approximation $\overline{\mathbf{H}} \approx_\beta \mathbf{H}$ in Algorithm (9). Moreover, Point 1 in Lemma 23 and Hessian comparison indicate $\mathbf{C}_i^2 = \widetilde{\mathbf{C}}(x_i, \nabla^2\phi_i(x_i)^2) \approx_{4\sigma} \nabla^2\phi_i(x_i)^{-1} \approx_{4\alpha} \nabla^2\phi_i(\overline{x}_i)^{-1}$. We then invoke Lemma 21 and Lemma 61 to obtain

$$\overline{\tau}_i \approx_\alpha \tau_i = w(\mathbf{C})_i \gtrsim \sigma_i(\mathbf{T}^{-\frac{1}{2}}\mathbf{C}\mathbf{A}) \approx_{O(\alpha+\sigma)} \sigma_i(\overline{\mathbf{T}}^{-\frac{1}{2}}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}).$$

Therefore we conclude that $\left\|\delta_{2,i}\right\|_2 \lesssim \left\|\mathbf{A}^\top x - b\right\|_{\mathbf{H}^{-1}}$ holds for any $i \in [m]$. For the τ-norm of δ_2, it holds that

$$\begin{aligned}
\left\|\delta_2\right\|_{2,\tau} = \left\|\delta_2\right\|_{\mathsf{diag}_{\mathrm{G}}(\tau)} &= \left\|\overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}(\mathbf{A}^\top x - b)\right\|_{\mathsf{diag}_{\mathrm{G}}(\tau)} \\
&\leq e^{\alpha}\left\|\mathbf{A}^\top x - b\right\|_{\overline{\mathbf{H}}^{-1}\mathbf{H}\overline{\mathbf{H}}^{-1}} \\
&\leq e^{\alpha+\beta}\left\|\mathbf{A}^\top x - b\right\|_{\mathbf{H}^{-1}}
\end{aligned}$$

due to $\tau \approx_\alpha \overline{\tau}$, and $\mathbf{H} \approx_\alpha \overline{\mathbf{H}}$. Combining the above bounds, we obtain

$$\left\|\nabla^2\Phi(x)^{\frac{1}{2}}\delta_2\right\|_{2,\tau+\infty} \leq 2\left(C_{\mathrm{norm}} + 1\right)\left\|\mathbf{A}^\top x - b\right\|_{\mathbf{H}^{-1}} \lesssim \alpha\beta.$$

where the second step is due to approximate feasibility, α-approximation, Corollary 25. We proceed to bound $g - \delta_1$. It simply holds that

$$\left\|\nabla^2\Phi(x)^{\frac{1}{2}}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}(g-\delta_1)\right\|_{2,\tau+\infty} \le e^{2\alpha}\left\|g-\delta_1\right\|_{2,\tau+\infty}.$$

In particular, the τ-norm can be controlled as

$$\begin{aligned}
\left\|g-\delta_1\right\|_{2,\tau} &\le e^{\alpha}\left\|\left(\mathbf{I}-\overline{\mathbf{T}}^{-\frac{1}{2}}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-\frac{1}{2}}\right)\overline{\mathbf{T}}^{\frac{1}{2}}g\right\|_2 \\
&\le e^{\alpha}\left\|\left(\mathbf{I}-\mathbf{P}\left(\overline{\mathbf{T}}^{-\frac{1}{2}}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\right)\right)\overline{\mathbf{T}}^{\frac{1}{2}}g\right\|_2 \\
&+ e^{\alpha}\left\|\overline{\mathbf{T}}^{-\frac{1}{2}}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\left(\overline{\mathbf{H}}^{-1}-\mathbf{H}^{-1}\right)\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-\frac{1}{2}}\overline{\mathbf{T}}^{\frac{1}{2}}g\right\|_2 \\
&\le e^{\alpha}\left\|\left(\mathbf{I}-\mathbf{P}\left(\overline{\mathbf{T}}^{-\frac{1}{2}}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\right)\right)\overline{\mathbf{T}}^{\frac{1}{2}}g\right\|_2 + 2\beta e^{\alpha}\left\|\mathbf{P}\left(\overline{\mathbf{T}}^{-\frac{1}{2}}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\right)\overline{\mathbf{T}}^{\frac{1}{2}}g\right\|_2 \\
&\le e^{\alpha}(1+2\beta)\left\|\overline{\mathbf{T}}^{\frac{1}{2}}g\right\|_2 \le \left(1+O(\alpha+\beta)\right)\left\|g\right\|_{2,\tau}.
\end{aligned}$$

where we use triangular inequality, blockwise spectral approximation $\overline{\mathbf{H}} \approx_\beta \mathbf{H}$ in (9), and $0 \preceq \mathbf{P}\left(\overline{\mathbf{T}}^{-\frac{1}{2}}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\right) \preceq \mathbf{I}$ (Lemma 55). For the ℓ_∞-norm, it holds that

$$\begin{aligned}
\left\|(\delta_1)_i\right\|_2 &= \left\|\mathbf{E}_i\overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g\right\|_2 \\
&\le \left\|\mathbf{E}_i\overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-\frac{1}{2}}\right\|_2\left\|\overline{\mathbf{T}}^{\frac{1}{2}}g\right\|_2 \\
&\le \|g\|_{2,\overline{\tau}}\cdot\left(\lambda_{\max}\left(\mathbf{E}_i\overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}\mathbf{H}\overline{\mathbf{H}}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-1}\mathbf{E}_i\right)\right)^{\frac{1}{2}} \\
&\le e^{\alpha+\beta}\|g\|_{2,\tau}\cdot\overline{\tau}_i^{-\frac{1}{2}}\left(\lambda_{\max}\left(\mathbf{E}_i\overline{\mathbf{T}}^{-\frac{1}{2}}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\mathbf{H}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-1}\mathbf{E}_i\right)\right)^{\frac{1}{2}} \\
&= e^{\alpha+\beta}\|g\|_{2,\tau}\cdot\overline{\tau}_i^{-\frac{1}{2}}\left(\operatorname{tr}\left(\mathbf{P}\left(\overline{\mathbf{T}}^{-\frac{1}{2}}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\right)_{ii}\right)\right)^{\frac{1}{2}} \\
&\le e^{\alpha+\beta}\|g\|_{2,\tau}
\end{aligned}$$

due to $\overline{\mathbf{H}} \approx_\beta \mathbf{H}$, $\overline{\tau} \approx_\alpha \tau$ and Lemma 61. This immediately implies $\|\delta_1\|_{2,\infty} \le 2\|g\|_{2,\tau}$. It then follows

$$\begin{aligned}
\left\|g-\delta_1\right\|_{2,\tau+\infty} &\le (1+O(\alpha+\beta))C_{\text{norm}}\left\|g\right\|_{2,\tau} + \left\|g\right\|_{2,\infty} + (1+2\alpha+2\beta)\left\|g\right\|_{2,\tau} \\
&\le (1+O(C_{\text{norm}}^{-1}+\alpha+\beta))\left\|g\right\|_{2,\tau+\infty}.
\end{aligned}$$

Because dual vector has unit length under the primal norm and $\tau \approx_\alpha \overline{\tau}$, we obtain the mixed norm of g

$$\left\|g\right\|_{2,\tau+\infty} = \beta\left\|\nabla_\mu^{\text{loc}}\Psi(\overline{\mu},\overline{x})^{\flat(\overline{\tau})}\right\|_{2,\tau+\infty} \le e^{\alpha}\beta\left\|\nabla_\mu^{\text{loc}}\Psi(\overline{\mu},\overline{x})^{\flat(\overline{\tau})}\right\|_{2,\overline{\tau}+\infty} \le (1+O(\alpha))\beta,$$

Collecting the above results, it holds that

$$\left\|\nabla^2\Phi(x)^{\frac{1}{2}}\mathbb{E}[\delta_x]\right\|_{2,\tau+\infty} \le (1+O(C_{\text{norm}}^{-1}+\alpha+\beta))\beta.$$

Bound on $\left\|\nabla^2\Phi(x)^{\frac{1}{2}}\delta_x\right\|_{2,\infty}$. We use triangular inequality to split the deterministic and stochastic part

$$\left\|\nabla^2\Phi(x)^{\frac{1}{2}}\delta_x\right\|_{2,\infty} \le e^{2\alpha}\left\|g-\mathbf{R}\delta_r\right\|_{2,\infty} \le e^{2\alpha}\left\|g-\delta_r\right\|_{2,\infty} + e^{2\alpha}\left\|\delta_r-\mathbf{R}\delta_r\right\|_{2,\infty}.$$

The deterministic part

$$\left\|(g-\delta_r)\right\|_{2,\infty} \le \left\|g\right\|_{2,\infty} + \left\|\delta_1\right\|_{2,\infty} + \left\|\delta_2\right\|_{2,\infty} \le \left\|g\right\|_{2,\infty} + 2\left\|g\right\|_{2,\tau} + \left\|\delta_2\right\|_{2,\infty}.$$

We apply the Maximum property in Definition 19 to bound the stochastic part

$$\left\|\delta_r - \mathbf{R}\delta_r\right\|_{2,\infty} \le \sqrt{\frac{d}{m}}\frac{\beta}{C_{\mathrm{var}}^2}\left\|\delta_r\right\|_2 \le \frac{\beta}{C_{\mathrm{var}}^2}\left(\left\|\delta_1\right\|_{2,\tau} + \left\|\delta_2\right\|_{2,\tau}\right).$$

As a consequence, we have

$$\begin{aligned}\left\|\nabla^2\Phi(\overline{x})^{\frac{1}{2}}\delta_x\right\|_{2,\infty} &\le e^{2\alpha}\left(\left\|g\right\|_{2,\infty} + 2\left\|g\right\|_{2,\tau} + \left\|\delta_2\right\|_{2,\infty} + \frac{\beta}{C_{\mathrm{var}}^2}\left(\left\|\delta_1\right\|_{2,\tau} + \left\|\delta_2\right\|_{2,\tau}\right)\right)\\ &\le (1 + O(\alpha+\beta+C_{\mathrm{norm}}^{-1}))\beta\end{aligned}$$

with the previously established norm bounds.

Bound on $\left\|\left(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\|_2]\right)_{i\in[m]}\right\|_{\tau+\infty}$. For the third claim, we compute the expected ℓ_2 norm of the i-th block

$$\mathbb{E}\left[\|\nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}}\delta_{x,i}\|_2\right] = \mathbb{E}\left[\|g_i - R_{ii}\delta_{r,i}\|_2\right] \le \|g_i\|_2 + \mathbb{E}\left[R_i\right]\|\delta_{r,i}\|_2 \le \|g_i\|_2 + \|\delta_{1,i}\|_2 + \|\delta_{2,i}\|_2$$

where we use the fact $\mathbf{R}$ is block-diagonal and the Expectation property of the sampling scheme in Definition 19. This gives

$$\left\|\left(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\|]\right)_{i\in[m]}\right\|_{\tau+\infty} \lesssim \left\|g\right\|_{2,\tau+\infty} + \left\|\delta_1\right\|_{2,\tau+\infty} + \left\|\delta_2\right\|_{2,\tau+\infty} \lesssim \beta$$

using the previous norm of g, δ_1, and δ_2.

Bound on $\left\|\left(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\|_2^2]\right)_{i\in[m]}\right\|_{\tau+\infty}$. Similar to the previous claim, we take expectation over the squared norm of the i-th block

$$\mathbb{E}\left[\left\|\nabla^2\phi_i(\overline{x})^{\frac{1}{2}}\delta_{x,i}\right\|_2^2\right] = \mathbb{E}\left[\left\|(g_i - R_i\delta_{r,i})\right\|_2^2\right] \lesssim \left\|g_i\right\|_2^2 + \mathbb{E}\left[\left\|R_i\delta_{r,i}\right\|_2^2\right].$$

In particular, we use bias-variance decomposition and the Variance property of $\mathbf{R}$ to estimate

$$\mathbb{E}\left[\|R_i\delta_{r,i}\|_2^2\right] = \|\delta_{r,i}\|_2^2 + \mathrm{Var}(R_i\|\delta_{r,i}\|_2) \le \|\delta_{r,i}\|_2^2 + \frac{\beta\left\|\delta_r\right\|_{2,\tau}}{C_{\mathrm{Var}}^2}\left\|\delta_{r,i}\right\|_2 \lesssim \beta\left\|\delta_{r,i}\right\|_2$$

where, in addition, is due to $\frac{d}{m}\|\delta_r\|_2^2 \le \|\delta_r\|_{2,\tau}^2 \lesssim \|\delta_1\|_{2,\tau}^2 + \|\delta_2\|_{2,\tau}^2 \lesssim \beta^2$. Continuing with previous bounds, it holds that

$$\|\delta_{r,i}\|_2^2 \lesssim \|\delta_{1,i}\|_2^2 + \|\delta_{2,i}\|_2^2 \lesssim \beta\left\|\delta_{1,i}\right\|_2 + \beta\left\|\delta_{2,i}\right\|_2.$$

Similarly, $\|g_i\|_2^2 \lesssim \beta\left\|g_i\right\|_2$. As a result of the above computation, we arrive at

$$\left\|\left(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\|_2^2]\right)_{i\in[m]}\right\|_{\tau+\infty} \lesssim \beta\left\|g\right\|_{2,\tau+\infty} + \beta\left\|\delta_1\right\|_{2,\tau+\infty} + \beta\left\|\delta_2\right\|_{2,\tau+\infty} \lesssim \beta^2$$

by the monotonicity of the mixed norm. □

Proof of Corollary 27. Since $\delta_s = \rho\overline{\mathbf{T}}\nabla^2\Phi(\overline{x})^{\frac{1}{2}}\delta_1$, it is deterministic and the claimed bound is directly from Hessian comparison, α-approximation, and also the norm bound of δ_1 in Lemma 26 $\square$

Proof of Lemma 29. As the first step, we decompose l_i into a stochastic and a deterministic term

$$l_i = \left\|\mathrm{sym}(\mathbf{C}_i^{-1}\boldsymbol{\Delta}_{C,i})\right\|_{\mathrm{op}} \lesssim \left\|\nabla^2\phi_i(x_i)\delta_{x,i}\right\| \lesssim \left\|g_i\right\|_2 + \left\|(\mathbf{R}\delta_r)_i\right\|_2$$

where we use the same argument in the proof of Lemma 28 in the first step. For the second step, we use the update rule of δ and Hessian comparison in Lemma 3. Point 2 in Lemma 23 and Lemma 26 indicate that $\mathbf{C}^2 \approx_{O(\beta)} (\mathbf{C}^{(\mathrm{new})})^2$. It is easy to notice $(\boldsymbol{\Delta}_C)^2 \preceq O(\beta^2)\mathbf{C}^2$. The identity $(\mathbf{C}^{(\mathrm{new})})^2 - \mathbf{C}^2 = \mathbf{C}\boldsymbol{\Delta}_C + \boldsymbol{\Delta}_C\mathbf{C} + (\boldsymbol{\Delta}_C)^2$ implies

$$-O(\beta)\mathbf{C}^2 \preceq \mathbf{C}\boldsymbol{\Delta}_C + \boldsymbol{\Delta}_C\mathbf{C} \preceq O(\beta)\mathbf{C}^2.$$

Consequently, $\mathbf{C}_t^2 = \mathbf{C}^2 + t(\mathbf{C}\boldsymbol{\Delta}_C + \boldsymbol{\Delta}_C\mathbf{C}) + t^2(\boldsymbol{\Delta}_C)^2 \approx_{O(\beta)} \mathbf{C}^2$ for every $t \in [0,1]$. Similar results also holds between τ_t and τ as a consequence of Lemma 21. This allows us to invoke Lemma 56 and to obtain

$$\mathbb{E}\big[\|l\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\big] \lesssim \left\|g\right\|_{2,\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t} + \mathbb{E}\big[\|\mathbf{R}\delta_r\|_{2,\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\big] \lesssim \left\|g\right\|_{\tau} + \mathbb{E}\big[\|\mathbf{R}\delta_r\|_{2,\mathbf{P}\circ_{\mathrm{G}}\mathbf{P}}\big]$$

In particular, we have $\left\|g\right\|_{\tau} \lesssim \beta/C_{\mathrm{norm}}$ from the definition of g. Moreover, we use the Covariance property and $\mathrm{Var}[R_i] \leq \tau_i^{-1}$ of the sampling scheme to bound the latter term through Jensen

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{R}\delta_r\|_{2,\mathbf{P}\circ_{\mathrm{G}}\mathbf{P}}^2\big] &\lesssim \sum_{i\in[m]} \mathbb{E}[R_i^2]\mathrm{tr}\big(\mathbf{P}_i^2\big)\,\|\delta_{r,i}\|_2^2 + 2\sum_{i\neq j}\mathbb{E}[R_iR_j]\mathrm{tr}\big(\mathbf{P}_{ij}\mathbf{P}_{ji}\big)\,\|\delta_{r,i}\|_2\,\|\delta_{r,j}\|_2 \\ &\lesssim \|\delta_r\|_{2,\tau}^2 + \|\delta_r\|_{2,\mathbf{P}\circ_{\mathrm{G}}\mathbf{P}}^2 \lesssim \|\delta_1\|_{\tau}^2 + \|\delta_2\|_{\tau}^2 \lesssim (\beta/C_{\mathrm{norm}})^2,\end{aligned}$$

in addition, due to $\mathbf{P} \preceq \mathbf{T}$ from Lemma 55, and norm bound for δ_1, δ_2 from Lemma 26. The desired result follows from combining the results. $\square$

B.2 Stability of τ

In this subsection, we will prove Lemma 30. To begin with, we provide a decomposition lemma that we use frequently in the subsection.

Lemma 62. *Given $\alpha > 0$, projection matrix $\mathbf{P}$, block leverage score matrix $\boldsymbol{\Sigma} = \mathsf{diag}(\mathrm{tr}_{\mathrm{G}}(\mathbf{P}))$, and some diagonal positive-definite matrix $\mathbf{T} \succeq \boldsymbol{\Sigma}$. Define $\boldsymbol{\Lambda} = \boldsymbol{\Sigma} - \mathbf{P}\circ_{\mathrm{G}}\mathbf{P}$, and also*

$$\mathbf{R} = (\mathbf{T} + \alpha\boldsymbol{\Lambda})^{-1}\mathbf{T}, \qquad \mathbf{D} = (\mathbf{T} + \alpha\boldsymbol{\Sigma})^{-1}\mathbf{T}, \qquad \mathbf{K} = \alpha\mathbf{R}\mathbf{T}^{-1}(\mathbf{P}\circ_{\mathrm{G}}\mathbf{P})\mathbf{D}.$$

Then it holds that $\mathbf{R} = \mathbf{D} + \mathbf{K}$. Moreover, the following properties are true:

- *$\mathbf{D}$ is diagonal and satisfies $2\mathbf{D}\mathbf{T}^{-1}\boldsymbol{\Sigma} \preceq 2(\alpha+1)^{-1}\mathbf{I}$;*
- *$\mathbf{R}$ is a contraction in both the ℓ_∞- and τ-norms, and $\mathbf{K}$ is a contraction in ℓ_∞-norm. Moreover, for any $x \in \mathbb{R}^m$, it holds that $\|\mathbf{K}x\|_\infty \leq \alpha\,|||x|||_{\mathbf{P}\circ_{\mathrm{G}}\mathbf{P}}$, $\|\mathbf{K}x\|_\tau \leq \alpha\,|||x|||_{\mathbf{P}\circ_{\mathrm{G}}\mathbf{P}}$.*

Proof. We first verify the decomposition. Since $\mathbf{\Lambda} = \mathbf{\Sigma} - \mathbf{P} \circ_{\mathrm{G}} \mathbf{P}$, the resolvent identity gives

$$\begin{aligned}\mathbf{R} &= (\mathbf{T} + \alpha\mathbf{\Sigma})^{-1}\mathbf{T} + \alpha(\mathbf{T} + \alpha\mathbf{\Lambda})^{-1}(\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})(\mathbf{T} + \alpha\mathbf{\Sigma})^{-1}\mathbf{T} \\ &= \mathbf{D} + \alpha\mathbf{R}\mathbf{T}^{-1}(\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})\mathbf{D} = \mathbf{D} + \mathbf{K}.\end{aligned}$$

The first property follows from $0 \preceq \mathbf{\Sigma} \preceq \mathbf{T}$, since $2\sigma_i/(\tau_i + \alpha\sigma_i) \leq 2/(\alpha + 1)$.

For the second property, recall from Lemma 55 that $0 \preceq \mathbf{P} \circ_{\mathrm{G}} \mathbf{P} \preceq \mathbf{\Sigma} \preceq \mathbf{T}$ and $(\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})\mathbf{1} = \mathbf{\Sigma}\mathbf{1}$. Consequently, $\mathbf{\Lambda} \succeq 0$ and $\mathbf{\Lambda}\mathbf{1} = 0$. Since $\mathbf{I} + \alpha\mathbf{T}^{-1}\mathbf{\Lambda}$ is a nonsingular M-matrix, its inverse $\mathbf{R}$ is entrywise nonnegative. Together with $\mathbf{R}\mathbf{1} = \mathbf{1}$, this gives the ℓ_∞-norm contraction. The τ-norm contraction follows from

$$\mathbf{T}^{\frac{1}{2}}\mathbf{R}\mathbf{T}^{-\frac{1}{2}} = (\mathbf{I} + \alpha\overline{\mathbf{\Lambda}})^{-1} \preceq \mathbf{I},$$

where $\overline{\mathbf{\Lambda}} = \mathbf{T}^{-\frac{1}{2}}\mathbf{\Lambda}\mathbf{T}^{-\frac{1}{2}}$. Since $\mathbf{K} = \alpha\mathbf{R}\mathbf{T}^{-1}(\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})\mathbf{D}$ is entrywise nonnegative and $\mathbf{K}\mathbf{1} = \mathbf{1} - \mathbf{D}\mathbf{1} \leq \mathbf{1}$, $\mathbf{K}$ is also a contraction in ℓ_∞-norm.

It remains to prove the other norm bounds for $\mathbf{K}$. We first obtain $(\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})\mathbf{T}^{-1}(\mathbf{P} \circ_{\mathrm{G}} \mathbf{P}) \preceq \mathbf{P} \circ_{\mathrm{G}} \mathbf{P}$ from $\mathbf{P} \circ_{\mathrm{G}} \mathbf{P} \preceq \mathbf{T}$. Consequently,

$$\big\|\mathbf{K}x\big\|_\tau \leq \alpha\big\|\mathbf{T}^{-1}(\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})\mathbf{D}x\big\|_\tau \leq \alpha\big\|\mathbf{D}x\big\|_{\mathbf{P} \circ_{\mathrm{G}} \mathbf{P}} \leq \alpha\big\||x|\big\|_{\mathbf{P} \circ_{\mathrm{G}} \mathbf{P}}$$

where we use the contraction of $\mathbf{R}$, the entrywise nonnegativity of $\mathbf{P} \circ_{\mathrm{G}} \mathbf{P}$, and diagonal $\mathbf{D} \preceq \mathbf{I}$. Similarly, we compute

$$\big\|\mathbf{K}x\big\|_\infty \leq \alpha \max_i \big|\big(\mathbf{T}^{-1}(\mathbf{P} \circ_{\mathrm{G}} \mathbf{P})\mathbf{D}x\big)_i\big| \leq \alpha\big\|\mathbf{D}x\big\|_{\mathbf{P} \circ_{\mathrm{G}} \mathbf{P}} \leq \alpha\big\||x|\big\|_{\mathbf{P} \circ_{\mathrm{G}} \mathbf{P}}$$

where, in addition, we use Cauchy-Schwarz and $\big(\mathrm{tr}(\mathbf{P}_{ii}^2)\big)^{\frac{1}{2}} \leq \mathrm{tr}(\mathbf{P}_{ii}) \leq \tau_i$ in the second inequality. □

We introduce a lemma to characterize the relative differential property for τ.

Lemma 63. *Suppose the setting and definitions in Lemma 30. Define $\mathbf{C}_t = \mathbf{C} + t\mathbf{\Delta}_C$, $\tau_t = w(\mathbf{C}_t)$, and $\delta_{\tau_t} = \mathrm{d}\tau_t/\mathrm{d}t$. Then the following properties hold:*

- $\big\|\mathbf{T}^{-1}\delta_{\tau_t}\big\|_\infty \lesssim \beta$ *w.h.p.;*
- $\big\|\mathbb{E}[(\mathbf{T}^{-1}\delta_{\tau_t})^2]\big\|_{\tau+\infty} \leq \beta^2$*;*
- $\mathbb{E}\big[\|\mathbf{T}^{-1}\,|\delta_{\tau_t}|\|_{\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t}\big] \lesssim \beta/C_{norm}$*.*

Proof. The variational characterization of the block Lewis weights gives $\nabla_w f(\tau_t, \mathbf{C}_t) = 0$ for every $t \in [0, 1]$. We take the time derivative on both sides and rearrange as

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\left[\nabla_w f(\tau_t, \mathbf{C}_t)\right] = \nabla_{ww} f(\tau_t, \mathbf{C}_t)\delta_{\tau_t} + \left[\frac{\mathrm{d}}{\mathrm{d}t}\nabla_w f(w, \mathbf{C}_t)\right]_{w=\tau_t}.$$

Plugging in Lemma 59 and Lemma 60, we obtain

$$\mathbf{T}_t^{-1}\delta_{\tau_t} = 2\left(\mathbf{T}_t - \left(1 - \frac{2}{p}\right)\mathbf{P}_t \circ_{\mathrm{G}} (\mathbf{I} - \mathbf{P}_t)\right)^{-1} \mathrm{tr}_{\mathrm{G}}\left(\mathbf{P}_t \mathrm{sym}(\mathbf{C}_t^{-1}\mathbf{\Delta}_C)(\mathbf{I} - \mathbf{P}_t)\right)$$

$$= 2(\mathbf{T}_t + \alpha\mathbf{\Lambda}_t)^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathrm{sym}(\mathbf{C}_t^{-1}\mathbf{\Delta}_C)\mathbf{P}_t - \mathbf{P}_t\mathrm{sym}(\mathbf{C}_t^{-1}\mathbf{\Delta}_C)\mathbf{P}_t\right), \tag{31}$$

where $\alpha = 2/p - 1$, $\mathbf{P}_t = \mathbf{P}\left(\mathbf{T}_t^{1/2-1/p}\mathbf{C}_t\mathbf{A}\right)$, $\mathbf{\Lambda}_t = \mathbf{P}_t \circ_{\mathrm{G}} (\mathbf{I} - \mathbf{P}_t)$. It is easy to see $\mathbf{C}_t^2 \approx_{O(\beta)} \mathbf{C}^2$ is true for any $t \in [0,1]$ as a consequence of Lemma 26. By Lemma 21, we have $\tau_t \approx_{O(\beta)} \tau$. We will frequently use this fact in this subsection without direct referencing. Also, define $l_{t,i} = \left\|\mathrm{sym}(\mathbf{C}_{t,i}^{-1}\mathbf{\Delta}_{C,i})\right\|_2$. Since $\mathbf{C}_i$, $\mathbf{\Delta}_{C,i}$ are symmetric and $n_i = O(1)$, the resolvent argument in the proof of Lemma 28 gives $l_{t,i} \approx_{O(1)} l_i$. Consequently, we obtain

$$\left\|l_t\right\|_\infty \lesssim \beta, \qquad \left\|\mathbb{E}[l_t^2]\right\|_{\tau+\infty} \lesssim \beta^2, \qquad \mathbb{E}\left[\|l_t\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\right] \lesssim \frac{\beta}{C_{\mathrm{norm}}}. \tag{32}$$

from Lemma 28 and Lemma 29.

Invoking Lemma 62, we decompose the RHS in (31) into the following form

$$\begin{aligned}\mathbf{T}_t^{-1}\delta_{\tau_t} =& 2\underbrace{\mathbf{D}_t\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathrm{sym}(\mathbf{C}_t^{-1}\mathbf{\Delta}_C)\mathbf{P}_t\right)}_{\text{(I)}} + 2\underbrace{\mathbf{K}_t\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathrm{sym}(\mathbf{C}_t^{-1}\mathbf{\Delta}_C)\mathbf{P}_t\right)}_{\text{(II)}} \\ &- 2\underbrace{\mathbf{R}_t\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{P}_t\mathrm{sym}(\mathbf{C}_t^{-1}\mathbf{\Delta}_C)\mathbf{P}_t\right)}_{\text{(III)}}.\end{aligned} \tag{33}$$

We now show the first property. For any $i \in [m]$, it holds that

$$\left|(\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathrm{sym}(\mathbf{C}_t^{-1}\mathbf{\Delta}_C)\mathbf{P}_t\right))_i\right| \le \tau_{t,i}^{-1}\left|\mathrm{tr}(\mathrm{sym}(\mathbf{C}_{t,i}^{-1}\mathbf{\Delta}_{C,i})\mathbf{P}_{t,i})\right| \le \left\|\mathrm{sym}(\mathbf{C}_{t,i}^{-1}\mathbf{\Delta}_{C,i})\right\| = l_{t,i}$$

due to $\mathrm{tr}(\mathbf{P}_{t,i}) \le \tau_{t,i}$, and the fact that $\mathrm{tr}(\mathbf{AB}) \le \mathrm{tr}(\mathbf{A})\left\|\mathbf{B}\right\|_2$ holds for any symmetric $\mathbf{B}$ and $\mathbf{A} \succeq 0$. Therefore the inequality $\left|\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathrm{sym}(\mathbf{C}_t^{-1}\mathbf{\Delta}_C)\mathbf{P}_t\right)\right| \le \mathbf{T}_t^{-1}\mathbf{\Sigma}_t l_t \le l_t$ holds in a coordinate-wise sense. In a similar fashion, it holds that $\left|\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{P}_t\mathrm{sym}(\mathbf{C}_t^{-1}\mathbf{\Delta}_C)\mathbf{P}_t\right)\right| \le \mathbf{T}_t^{-1}(\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t)l_t$. We compute

$$\begin{aligned}\left\|\mathbf{T}^{-1}\delta_{\tau_t}\right\|_\infty &\le \left\|\mathbf{D}_t\mathbf{T}_t^{-1}\mathbf{\Sigma}_t l_t\right\|_\infty + \left\|\mathbf{K}_t\mathbf{T}_t^{-1}\mathbf{\Sigma}_t l_t\right\|_\infty + \left\|\mathbf{R}_t\mathbf{T}_t^{-1}(\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t)l_t\right\|_\infty \\ &\lesssim \left\|l_t\right\|_\infty + \left\|\mathbf{T}_t^{-1}\mathbf{\Sigma}_t l_t\right\|_\infty + \left\|\mathbf{T}_t^{-1}(\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t)l_t\right\|_\infty \lesssim \left\|l_t\right\|_\infty \lesssim \beta.\end{aligned}$$

In the first line we use diagonal bound on $\mathbf{D}_t$, and ℓ_∞-norm contractions of $\mathbf{K}_t$ and $\mathbf{R}_t$ from Lemma 62 . The second line requires $\mathbf{\Sigma}_t \preceq \mathbf{T}_t$ as a result of regularization, (iv) from Lemma 55, and also the bounds in (32).

For the second property, we apply the above coordinate-wise bound and obtain the upper bound for (I)

$$\left\|\mathbb{E}[(\mathrm{I})^2]\right\|_{\tau+\infty} \lesssim \left\|\mathbb{E}[l_t^2]\right\|_{\tau+\infty} \lesssim \beta^2$$

where the last step is due to (32). For (II), it holds that

$$\begin{aligned}\left\|\mathbb{E}[(\mathrm{II})^2]\right\|_{\tau+\infty} &\lesssim C_{\mathrm{norm}}\mathbb{E}\left[\|(\mathrm{II})^2\|_\tau\right] + \mathbb{E}\left[\|(\mathrm{II})\|_\infty^2\right] \\ &\lesssim \beta C_{\mathrm{norm}}\mathbb{E}\left[\|(\mathrm{II})\|_\tau\right] + \beta^2 \lesssim C_{\mathrm{norm}}\beta\mathbb{E}\left[\|l_t\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\right] + \beta^2 \lesssim \beta^2.\end{aligned}$$

Similarly, we use the τ_t-norm contraction of $\mathbf{R}_t$ to bound (III).Combining the above, we obtain $\left\|\mathbb{E}[(\mathbf{T}_t^{-1}\delta_{\tau_t})^2]\right\|_{\tau+\infty} \lesssim \beta^2$. For the third property, we compute

$$\mathbb{E}\left[\||\mathbf{T}^{-1}\delta_{\tau_t}|\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\right] \lesssim \mathbb{E}\left[\,\||(\mathrm{I})|\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t} + \|(\mathrm{II})\|_{\tau_t} + \|(\mathrm{III})\|_{\tau_t}\,\right] \lesssim \mathbb{E}\left[\,\|l_t\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\,\right] \lesssim \beta/C_{\mathrm{norm}}.$$

where we use entrywise nonnegativity of $\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t$, $\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t \preceq \mathbf{T}_t$ from Lemma 55, Point 2 of Lemma 62, and previous bounds. □

As the next step, we control the first-order Taylor remainder of between τ and $\tau^{(\text{new})}$.

Lemma 64. *Suppose the setting and definitions in Lemma 30. Then the following inequality holds*

$$\left\|\mathbf{T}^{-1}\left(\mathbb{E}[\delta_\tau]-\mathbf{J}_C[\mathbb{E}\boldsymbol{\Delta}_C]\right)\right\|_{\tau+\infty}\lesssim\beta^2$$

where $\mathbf{J}_C$ *is the Jacobian tensor at* $\mathbf{C}$.

Proof. The difference can be expressed through the following integral over second-order time derivative of τ_t

$$\mathbf{T}^{-1}\left(\mathbb{E}[\delta_\tau]-\mathbf{J}_{\mathbf{C}}[\mathbb{E}\boldsymbol{\Delta}_C]\right)=\int_0^1(1-t)\mathbb{E}\left[\mathbf{T}^{-1}\frac{\mathrm{d}^2\tau_t}{\mathrm{d}t^2}\right]\mathrm{d}t. \tag{34}$$

where we obtain the expression of the first-order derivative in the same way in Lemma 63

$$\frac{\mathrm{d}\tau_t}{\mathrm{d}t}=\mathbf{T}_t\left(\mathbf{T}_t-\left(1-\frac{2}{p}\right)\mathbf{P}_t\circ_{\mathrm{G}}(\mathbf{I}_n-\mathbf{P}_t)\right)^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{P}_t\frac{\boldsymbol{\Delta}_C}{\mathbf{C}_t}(\mathbf{I}_n-\mathbf{P}_t)\right)$$

We define the following matrices

$$\mathbf{R}_t=(\mathbf{T}_t+\alpha\boldsymbol{\Lambda}_t)^{-1}\mathbf{T}_t,\qquad u_t=\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}(\mathbf{S}_t\mathbf{P}_t-\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t)$$

where $\alpha=2/p-1$. We also define auxiliary notations like $\mathbf{S}_t=\mathrm{sym}(\mathbf{C}_t^{-1}\boldsymbol{\Delta}_C)$, $h_t=\mathbf{T}_t^{-1}\delta_{\tau_t}$. As a consequence, it holds that

$$\mathbf{T}_t^{-1}\delta_{\tau_t}=2\mathbf{R}_tu_t,\qquad\mathbf{T}_t^{-1}\frac{\mathrm{d}^2\tau_t}{\mathrm{d}t^2}=h_t^2+2\dot{\mathbf{R}}_tu_t+2\mathbf{R}_t\dot{u}_t.$$

Moreover, let $\mathbf{L}_t=\mathbf{T}_t^{1/2-1/p}\mathbf{C}_t$. We express the time derivative of $\mathbf{L}_t$ and $\mathbf{P}_t=\mathbf{P}(\mathbf{L}_t\mathbf{A})$ as

$$\dot{\mathbf{L}}_t\mathbf{L}_t^{-1}=\boldsymbol{\Delta}_C\mathbf{C}_t^{-1}-\frac{\alpha}{2}\mathbf{H}_t$$

and

$$\dot{\mathbf{P}}_t=\boldsymbol{\Delta}_C\mathbf{C}_t^{-1}\mathbf{P}_t+\mathbf{P}_t\mathbf{C}_t^{-1}\boldsymbol{\Delta}_C-2\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t-\frac{\alpha}{2}\big(\mathbf{H}_t\mathbf{P}_t+\mathbf{P}_t\mathbf{H}_t-2\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\big).$$

It requires to expand the second-order time derivative before we bound the individual terms.

Expansion of $\mathbf{R}_t$. We first analyze and take time derivative of $\mathbf{R}_t=(\mathbf{T}_t+\alpha\boldsymbol{\Lambda}_t)^{-1}\mathbf{T}_t$

$$\dot{\mathbf{R}}_t=\mathbf{R}_t\mathbf{H}_t-(1+\alpha)\mathbf{R}_t\mathbf{H}_t\mathbf{R}_t+\alpha\mathbf{R}_t\mathbf{T}_t^{-1}\frac{d}{dt}(\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t)\mathbf{R}_t.$$

due to $\dot{\boldsymbol{\Sigma}}_t=\mathbf{T}_t\mathbf{H}_t$ and $\dot{\boldsymbol{\Lambda}}_t=\mathbf{T}_t\mathbf{H}_t-\frac{d}{dt}(\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t)$. Using $2\mathbf{R}_tu_t=h_t$, this implies

$$2\dot{\mathbf{R}}_tu_t=2\mathbf{R}_t\mathbf{H}_tu_t-(1+\alpha)\mathbf{R}_th_t^2+\alpha\mathbf{R}_t\mathbf{T}_t^{-1}\left(\frac{d}{dt}(\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t)\right)h_t.$$

By the definition of the block Hadamard product, the last term is expanded as

$$\begin{aligned}\left(\frac{d}{dt}(\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t)\right)h_t=2\mathrm{tr}_{\mathrm{G}}(\dot{\mathbf{P}}_t\mathbf{H}_t\mathbf{P}_t)=&-\alpha\mathrm{tr}_{\mathrm{G}}\big(\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t+\mathbf{P}_t\mathbf{H}_t^2\mathbf{P}_t-2\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\big)\\&+2\mathrm{tr}_{\mathrm{G}}\big(\mathbf{S}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t+\mathbf{P}_t\mathbf{S}_t\mathbf{H}_t\mathbf{P}_t-2\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\big).\end{aligned}$$

where the asymmetric parts cancel after taking the block traces.

Expansion of u_t. We next take time derivative of u_t as

$$\dot{u}_t = -\mathbf{H}_t u_t + \mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\dot{\mathbf{S}}_t\mathbf{P}_t + \mathbf{S}_t\dot{\mathbf{P}}_t - \dot{\mathbf{P}}_t\mathbf{S}_t\mathbf{P}_t - \mathbf{P}_t\dot{\mathbf{S}}_t\mathbf{P}_t - \mathbf{P}_t\mathbf{S}_t\dot{\mathbf{P}}_t\right).$$

In particular, the asymmetric terms $\mathbf{B}_t = \mathbf{\Delta}_C\mathbf{C}_t^{-1}$ satisfy the exact identity

$$\dot{\mathbf{S}}_t + \mathbf{S}_t\mathbf{B}_t + \mathbf{B}_t^\top\mathbf{S}_t = \mathbf{B}_t^\top\mathbf{B}_t.$$

We substitute the expression for $\dot{\mathbf{P}}_t$ and replace the majority of terms where $\mathbf{B}_t$ and $\mathbf{B}_t^\top$ arises by regrouping. As a consequence, it holds that

$$\begin{aligned}\dot{u}_t =& -(1+\alpha)\mathbf{H}_t u_t + \alpha\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{S}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t + \mathbf{P}_t\mathbf{H}_t\mathbf{S}_t\mathbf{P}_t - 2\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\right)\\ &+ \mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{B}_t^\top\mathbf{B}_t\mathbf{P}_t - \mathbf{P}_t\mathbf{B}_t^\top\mathbf{B}_t\mathbf{P}_t\right) - 4\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{S}_t\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t - \mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\right).\end{aligned}$$

We now combine the expressions for $\dot{\mathbf{R}}_t$ and $\dot{u}_t$ to obtain

$$\begin{aligned}\mathbb{E}\left[\mathbf{T}^{-1}\ddot{\tau}_t\right] =& \underbrace{\mathbb{E}\left[\mathbf{T}^{-1}\mathbf{T}_t h_t^2\right]}_{\text{(I)}} - (1+\alpha)\underbrace{\mathbb{E}\left[\mathbf{T}^{-1}\mathbf{T}_t\mathbf{R}_t h_t^2\right]}_{\text{(II)}} - 2\alpha\underbrace{\mathbb{E}\left[\mathbf{T}^{-1}\mathbf{T}_t\mathbf{R}_t\mathbf{H}_t u_t\right]}_{\text{(III)}}\\ &- \alpha^2\underbrace{\mathbb{E}\left[\mathbf{T}^{-1}\mathbf{T}_t\mathbf{R}_t\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t + \mathbf{P}_t\mathbf{H}_t^2\mathbf{P}_t - 2\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\right)\right]}_{\text{(IV)}}\\ &+ 4\alpha\underbrace{\mathbb{E}\left[\mathbf{T}^{-1}\mathbf{T}_t\mathbf{R}_t\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{S}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t + \mathbf{P}_t\mathbf{S}_t\mathbf{H}_t\mathbf{P}_t - 2\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\right)\right]}_{\text{(V)}}\\ &- 8\underbrace{\mathbb{E}\left[\mathbf{T}^{-1}\mathbf{T}_t\mathbf{R}_t\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{S}_t\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t - \mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\right)\right]}_{\text{(VI)}}\\ &+ 2\underbrace{\mathbb{E}\left[\mathbf{T}^{-1}\mathbf{T}_t\mathbf{R}_t\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{B}_t^\top\mathbf{B}_t\mathbf{P}_t - \mathbf{P}_t\mathbf{B}_t^\top\mathbf{B}_t\mathbf{P}_t\right)\right]}_{\text{(VII)}}. \end{aligned} \tag{35}$$

We proceed to bound the terms in each term in (35). We will use the following facts without further referencing: $\mathbf{T}_t \approx_{O(\beta)} \mathbf{T}$. Let $l_{t,i} = \left\|\mathbf{S}_{t,i}\right\|_2$. By the established bounds, Lemma 28, Lemma 29 and Lemma 63, we have

$$\left\|h_t\right\|_\infty + \left\|l_t\right\|_\infty \lesssim \beta, \qquad \left\|\mathbb{E}[h_t^2 + l_t^2]\right\|_{\tau+\infty} \lesssim \beta^2, \qquad \mathbb{E}\left\|\,|h_t| + l_t\right\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t} \lesssim \frac{\beta}{C_{\mathrm{norm}}}.$$

Bounding simple terms. For the term (I), the following holds immediately

$$\left\|\mathbb{E}[\mathbf{T}^{-1}\mathbf{T}_t h_t^2]\right\|_{\tau+\infty} \lesssim \left\|\mathbb{E}[(\mathbf{T}_t^{-1}\delta_{\tau_t})^2]\right\|_{\tau+\infty} \lesssim \beta^2.$$

For the term (II), we invoke Lemma 62 and decompose $\mathbf{R}_t = \mathbf{D}_t + \mathbf{K}_t$. As a consequence, it holds that

$$\begin{aligned}\left\|\mathbb{E}[\mathbf{T}^{-1}\mathbf{T}_t\mathbf{R}_t h_t^2]\right\|_{\tau+\infty} &\lesssim \left\|\mathbb{E}[\mathbf{D}_t h_t^2]\right\|_{\tau+\infty} + \left\|\mathbb{E}[\mathbf{K}_t h_t^2]\right\|_\infty + C_{\mathrm{norm}}\left\|\mathbb{E}[\mathbf{K}_t h_t^2]\right\|_\tau\\ &\lesssim \left\|\mathbb{E}[(\mathbf{T}^{-1}\delta_{\tau_t})^2]\right\|_{\tau+\infty} + \mathbb{E}[\|\mathbf{T}^{-1}\delta_{\tau_t}\|_\infty^2] + \alpha\beta C_{\mathrm{norm}}\mathbb{E}[\||\mathbf{T}^{-1}\delta_{\tau_t}|\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}] \lesssim \beta^2\end{aligned}$$

where we use the facts from Point 2 in Lemma 62: $\mathbf{D}_t$ is diagonal with $0 \preceq \mathbf{D}_t \preceq \mathbf{I}$, $\mathbf{K}_t$ induces a contraction for ℓ_∞-norm.

For the term (III), we write down the expansion of $\mathbf{H}_t u_t$

$$\mathbf{H}_t u_t = \mathbf{H}_t \mathbf{T}_t^{-1} \mathrm{tr}_{\mathrm{G}} \left(\mathbf{S}_t \mathbf{P}_t\right) - \mathbf{H}_t \mathbf{T}_t^{-1} \mathrm{tr}_{\mathrm{G}} \left(\mathbf{P}_t \mathbf{S}_t \mathbf{P}_t\right). \tag{36}$$

For the first part in (36), the following inequality holds in coordinates

$$\left|\mathbf{T}_t^{-1} \mathrm{tr}_{\mathrm{G}} \left(\mathbf{S}_t \mathbf{P}_t\right)\right| \le \left|\mathbf{T}_t^{-1} \mathbf{\Sigma}_t\right| \le l_t$$

We use the decomposition in Lemma 62 and the coordinatewise inequality $2\left|\mathbf{H}_t\right| l_t \le h_t^2 + l_t^2$. It holds that

$$\begin{aligned}
\left\|\mathbb{E}\left[\mathbf{T}^{-1} \mathbf{T}_t \mathbf{R}_t \mathbf{H}_t \mathbf{T}_t^{-1} \mathrm{tr}_{\mathrm{G}} \left(\mathbf{S}_t \mathbf{P}_t\right)\right]\right\|_{\tau+\infty} &\lesssim \left\|\mathbb{E}[|\mathbf{H}_t| \, l_t]\right\|_{\tau+\infty} + \mathbb{E}[\|\mathbf{H}_t l_t\|_\infty] + \alpha C_{\mathrm{norm}} \mathbb{E}\left[\||\mathbf{H}_t| l_t\|_{\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t}\right] \\
&\lesssim \left\|\mathbb{E}[l_t^2]\right\|_{\tau+\infty} + \left\|\mathbb{E}[(\mathbf{T}_t^{-1} \delta_{\tau_t})^2]\right\|_{\tau+\infty} + \alpha C_{\mathrm{norm}} \beta \mathbb{E} \left\|l_t\right\|_{\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t} \\
&\lesssim \beta^2.
\end{aligned}$$

For the second part in (36), we compute coordinatewisely

$$\left|\mathbf{T}_t^{-1} \mathrm{tr}_{\mathrm{G}} \left(\mathbf{P}_t \mathbf{S}_t \mathbf{P}_t\right)\right| \le \mathbf{T}_t^{-1} \left(\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t\right) l_t, \qquad \left\|\mathbf{T}_t^{-1} \left(\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t\right) l_t\right\|_{\tau_t} \le \left\|l_t\right\|_{\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t},$$

due to $0 \preceq \mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t \preceq \mathbf{T}_t$ (Lemma 55). Lemma 62 indicates that $\mathbf{R}_t$ induces a contraction for both norms. Consequently, we obtain

$$\begin{aligned}
\left\|\mathbb{E}\left[\mathbf{T}^{-1} \mathbf{T}_t \mathbf{R}_t \mathbf{H}_t \mathbf{T}_t^{-1} \mathrm{tr}_{\mathrm{G}} \left(\mathbf{P}_t \mathbf{S}_t \mathbf{P}_t\right)\right]\right\|_{\tau+\infty} &\lesssim \mathbb{E}\left[\|\mathbf{H}_t \mathbf{T}_t^{-1} \mathrm{tr}_{\mathrm{G}} \left(\mathbf{P}_t \mathbf{S}_t \mathbf{P}_t\right)\|_{\tau+\infty}\right] \\
&\lesssim \beta \mathbb{E}\left[\|\mathbf{T}_t^{-1} \left(\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t\right) l_t\|_\infty\right] + \beta C_{\mathrm{norm}} \mathbb{E}\left[\|\mathbf{T}_t^{-1} \left(\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t\right) l_t\|_\tau\right] \\
&\lesssim \beta \mathbb{E}\left[\|l_t\|_\infty\right] + \beta C_{\mathrm{norm}} \mathbb{E}\left[\|l_t\|_{\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t}\right] \lesssim \beta^2.
\end{aligned}$$

also due to (iv) in Lemma 55.

Bounding terms. We next bound the term (IV). It requires to bound everything in its parenthesis coordinatewisely. We compute

$$\left|\mathrm{tr}\left(\left(\mathbf{H}_t \mathbf{P}_t \mathbf{H}_t \mathbf{P}_t\right)_{ii}\right)\right| \le |h_{t,i}| \sum_j |h_{t,j}| \, \|\mathbf{P}_{t,ij}\|_{\mathrm{F}}^2 \le \|h_t\|_\infty \big(\left(\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t\right) |h_t| \big)_i.$$

This suggests $\mathrm{tr}\left(\mathbf{H}_t \mathbf{P}_t \mathbf{H}_t \mathbf{P}_t\right) \lesssim \beta \left(\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t\right) |h_t|$. Also, it holds that $0 \preceq \mathbf{P}_t \mathbf{H}_t \mathbf{P}_t \mathbf{H}_t \mathbf{P}_t = \left(\mathbf{P}_t \mathbf{H}_t \mathbf{P}_t\right)^2 \preceq \mathbf{P}_t \mathbf{H}_t^2 \mathbf{P}_t$, as $\mathbf{P}_t$ is a projection matrix. As a consequence, we have

$$0 \le \mathrm{tr}_{\mathrm{G}}(\mathbf{P}_t \mathbf{H}_t \mathbf{P}_t \mathbf{H}_t \mathbf{P}_t) \le \mathrm{tr}_{\mathrm{G}}(\mathbf{P}_t \mathbf{H}_t^2 \mathbf{P}_t) = \left(\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t\right) h_t^2.$$

Using $0 \preceq \mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t \preceq \mathbf{T}_t$ and its entrywise nonnegativity, we obtain

$$\left\|\mathbf{T}_t^{-1} \mathrm{tr}_{\mathrm{G}}(\mathbf{H}_t \mathbf{P}_t \mathbf{H}_t \mathbf{P}_t + \mathbf{P}_t \mathbf{H}_t^2 \mathbf{P}_t - 2 \mathbf{P}_t \mathbf{H}_t \mathbf{P}_t \mathbf{H}_t \mathbf{P}_t)\right\|_\infty \lesssim \beta^2,$$

and the corresponding τ-norm is bounded by

$$\left\|\mathbf{T}_t^{-1} \mathrm{tr}_{\mathrm{G}}(\mathbf{H}_t \mathbf{P}_t \mathbf{H}_t \mathbf{P}_t + \mathbf{P}_t \mathbf{H}_t^2 \mathbf{P}_t - 2 \mathbf{P}_t \mathbf{H}_t \mathbf{P}_t \mathbf{H}_t \mathbf{P}_t)\right\|_\tau \lesssim \left\|h_t\right\|_\infty \left\||h_t|\right\|_{\mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t}.$$

Combining the above estimates and contraction property of $\mathbf{R}_t$, we obtain the following

$$\begin{aligned}\big\|(\mathrm{IV})\big\|_{\tau+\infty} &\lesssim \mathbb{E}\big[\,\big\|\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t+\mathbf{P}_t\mathbf{H}_t^2\mathbf{P}_t-2\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\right)\big\|_{\infty}\,\big]\\ &\quad+C_{\mathrm{norm}}\mathbb{E}\big[\,\big\|\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t+\mathbf{P}_t\mathbf{H}_t^2\mathbf{P}_t-2\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\right)\big\|_{\tau}\,\big]\\ &\lesssim \beta^2+\beta C_{\mathrm{norm}}\mathbb{E}\big[\|\,|h_t|\,\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\big]=\beta^2+\beta C_{\mathrm{norm}}\mathbb{E}\big[\|\,|\mathbf{T}_t^{-1}\delta_{\tau_t}|\,\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\big]\lesssim\beta^2.\end{aligned}$$

We turn to the mixed term (V). The following holds coordinatewise,

$$\left|\mathrm{tr}_{\mathrm{G}}\left(\mathbf{S}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\right)\right|_i\le\sum_j|h_{t,j}|\left|\mathrm{tr}\left(\mathbf{S}_{t,i}\mathbf{P}_{t,ij}\mathbf{P}_{t,ji}\right)\right|\le l_{t,i}\sum_j|h_{t,j}|\left\|\mathbf{P}_{t,ij}\right\|_{\mathrm{F}}^2=l_{t,i}\left(\left(\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t\right)|\mathbf{H}_t|\right)_i.$$

Similarly, we have $\left|\mathrm{tr}_{\mathrm{G}}\left(\mathbf{P}_t\mathbf{S}_t\mathbf{H}_t\mathbf{P}_t\right)\right|\le\left(\left(\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t\right)l_t\,|h_t|\right)_i$. For the last term in (V)'s parenthesis, it holds that

$$\left|\mathrm{tr}\left(\left(\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t\right)_{ii}\right)\right|\le\frac{1}{2}\mathrm{tr}\left(\left(\mathbf{P}_t\mathbf{S}_t^2\mathbf{P}_t+\mathbf{P}_t\mathbf{H}_t^2\mathbf{P}_t\right)_{ii}\right)\le\frac{1}{2}\left(\left(\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t\right)\left(l_t^2+h_t^2\right)\right)_i.$$

where we use Cauchy-Schwarz. Combining the above estimates and the contraction property of $\mathbf{R}_t$, we obtain the following

$$\begin{aligned}\big\|(\mathrm{V})\big\|_{\tau+\infty} &\lesssim \mathbb{E}\big[\|\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}(\mathbf{S}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t+\mathbf{P}_t\mathbf{S}_t\mathbf{H}_t\mathbf{P}_t-2\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t)\|_{\infty}\big]\\ &\quad+C_{\mathrm{norm}}\mathbb{E}\big[\|\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}(\mathbf{S}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t+\mathbf{P}_t\mathbf{S}_t\mathbf{H}_t\mathbf{P}_t-2\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\mathbf{H}_t\mathbf{P}_t)\|_{\tau_t}\big]\\ &\lesssim \beta^2+\beta C_{\mathrm{norm}}\mathbb{E}\big[\|l_t\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}+\big\||h_t|\big\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\big]\\ &=\beta^2+\beta C_{\mathrm{norm}}\mathbb{E}\big[\|l_t\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}+\||\mathbf{T}_t^{-1}\delta_{\tau_t}|\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\big]\lesssim\beta^2.\end{aligned}$$

We now bound (VI). It holds that

$$\left|\mathrm{tr}_{\mathrm{G}}(\mathbf{S}_t\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t)\right|_i\le\sum_j\left|\mathrm{tr}(\mathbf{S}_{t,i}\mathbf{P}_{t,ij}\mathbf{S}_{t,j}\mathbf{P}_{t,ji})\right|\le l_{t,i}\sum_j l_{t,j}\|\mathbf{P}_{t,ij}\|_{\mathrm{F}}^2=l_{t,i}((\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t)l_t)_i.$$

Also, since $\mathbf{P}_t\mathbf{S}_t^2\mathbf{P}_t-(\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t)^2\succeq 0$, we can bound as

$$0\le\mathrm{tr}_{\mathrm{G}}\left(\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\right)_i\le\mathrm{tr}\left(\left(\mathbf{P}_t\mathbf{S}_t^2\mathbf{P}_t\right)_{ii}\right)\le\sum_j l_{t,j}^2\left\|\mathbf{P}_{t,ij}\right\|_{\mathrm{F}}^2=\left(\left(\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t\right)l_t^2\right)_i.$$

We use the decomposition lemma and the above bounds to obtain

$$\begin{aligned}\big\|(\mathrm{VI})\big\|_{\tau+\infty} &\lesssim \mathbb{E}\big[\|\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}(\mathbf{S}_t\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t-\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t)\|_{\infty}\big]\\ &\quad+C_{\mathrm{norm}}\mathbb{E}\big[\|\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}(\mathbf{S}_t\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t-\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t\mathbf{S}_t\mathbf{P}_t)\|_{\tau}\big]\\ &\lesssim \beta^2+\beta C_{\mathrm{norm}}\mathbb{E}\big[\|l_t\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\big]\lesssim\beta^2.\end{aligned}$$

Bounding the asymmetric. Finally, we estimate the upper bound of the term (VII), where asymmetric term $\mathbf{B}_t=\boldsymbol{\Delta}_C\mathbf{C}_t^{-1}$ arises. We use Lemma 22 to obtain

$$\|\mathbf{B}_{t,i}\|_{\mathrm{op}}^2\le 2n_i\|\mathrm{sym}(\boldsymbol{\Delta}_C\mathbf{C}_t^{-1})\|_{\mathrm{op}}=2n_i l_{t,i}^2.$$

Since $n_i=O(1)$, it suggests that $0\le\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{L}_t^{\top}\mathbf{L}_t\mathbf{P}_t\right)\lesssim l_{t,i}^2$. Similarly, we have

$$0\le\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}(\mathbf{P}_t\mathbf{L}_t^{\top}\mathbf{L}_t\mathbf{P}_t)\lesssim\mathbf{T}_t^{-1}(\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t)l_t^2.$$

We now use the decomposition $\mathbf{R}_t = \mathbf{D}_t + \mathbf{K}_t$ in Lemma 62 again and compute

$$\begin{aligned}\|(\mathrm{VII})\|_{\tau+\infty} &\lesssim 2\big\|\mathbb{E}\big[\mathbf{D}_t\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}(\mathbf{B}_t^\top\mathbf{B}_t\mathbf{P}_t)\big]\big\|_{\tau+\infty} + \big\|\mathbb{E}[\mathbf{K}_t\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}(\mathbf{B}_t^\top\mathbf{B}_t\mathbf{P}_t)]\big\|_{\tau+\infty} \\ &\qquad + \big\|\mathbb{E}[\mathbf{R}_t\mathbf{T}_t^{-1}\mathrm{tr}_{\mathrm{G}}(\mathbf{P}_t\mathbf{B}_t^\top\mathbf{B}_t\mathbf{P}_t)]\big\|_{\tau+\infty} \\ &\lesssim \big\|\mathbb{E}[l_t^2]\big\|_{\tau+\infty} + \mathbb{E}\big[\|l_t^2\|_\infty\big] + C_{\mathrm{norm}}\mathbb{E}\big[\|l_t^2\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\big] \\ &\quad + \mathbb{E}\big[\|\mathbf{T}_t^{-1}(\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t)l_t^2\|_\infty\big] + C_{\mathrm{norm}}\mathbb{E}\big[\|\mathbf{T}_t^{-1}(\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t)l_t^2\|_{\tau_t}\big] \\ &\lesssim \beta^2 + C_{\mathrm{norm}}\mathbb{E}\big[\|l_t^2\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\big] \lesssim \beta^2 + \beta C_{\mathrm{norm}}\mathbb{E}\big[\|l_t\|_{\mathbf{P}_t\circ_{\mathrm{G}}\mathbf{P}_t}\big] \lesssim \beta^2.\end{aligned}$$

where we use $0 \preceq \mathbf{P}_t \circ_{\mathrm{G}} \mathbf{P}_t \preceq \mathbf{T}_t$, contraction properties from Lemma 62.

Combining the bounds for (I)–(VII), we conclude that $\big\|\mathbb{E}[\mathbf{T}^{-1}\ddot{\tau}_t]\big\|_{\tau+\infty} \lesssim \beta^2$ uniformly over $t \in [0,1]$. Substituting this estimate into the Taylor remainder identity completes the proof. □

We build upon Lemma 63 and Lemma 64 and obtain the following sharp bound for the expected change.

Lemma 65. *Suppose the setting and the definitions in Lemma 30. Then the following property holds*

$$\big\|\mathbf{T}^{-1}\mathbb{E}[\delta_\tau]\big\|_{\tau+\infty} \le \sqrt{p/2-p}\cdot\beta\beta + O(C_{norm}^{-1} + \alpha + \beta)\beta.$$

Proof. By the triangular inequality and the error bound in Lemma 64, we compute

$$\begin{aligned}\big\|\mathbf{T}^{-1}\mathbb{E}\left[\delta_\tau\right]\big\|_{\tau+\infty} &\le \big\|\mathbf{T}^{-1}\mathbf{J}_{\mathbf{C}}[\mathbb{E}\left[\boldsymbol{\Delta}_C\right]]\big\|_{\tau+\infty} + \big\|\mathbb{E}\left[\mathbf{T}^{-1}\left(\delta_\tau - \mathbf{J}_{\mathbf{C}}[\boldsymbol{\Delta}_C]\right)\right]\big\|_{\tau+\infty} \\ &\le \big\|\mathbf{T}^{-1}\mathbf{J}_{\mathbf{C}}[\mathbb{E}\left[\boldsymbol{\Delta}_C\right]]\big\|_{\tau+\infty} + O(\beta^2).\end{aligned}$$

It remains to bound the Jacobian term. We recall from the proof of Lemma 64 the following expression of the Jacobian term under expectation

$$\mathbf{T}^{-1}\mathbf{J}_{\mathbf{C}}[\mathbb{E}\boldsymbol{\Delta}_C] = 2\left(\mathbf{T}+\alpha\boldsymbol{\Lambda}\right)^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbb{E}\left[\mathrm{sym}\left(\mathbf{C}^{-1}\boldsymbol{\Delta}_C\right)\right]\mathbf{P} - \mathbf{P}\mathbb{E}\left[\mathrm{sym}\left(\mathbf{C}^{-1}\boldsymbol{\Delta}_C\right)\right]\mathbf{P}\right).$$

where $\alpha = 2/p - 1$. We first bound its τ-norm. Now denote $y = \left(\mathbf{T}+\alpha\boldsymbol{\Lambda}\right)^{-1}\mathrm{tr}_{\mathrm{G}}\left(\mathbf{SP} - \mathbf{PSP}\right)$, where $\mathbf{S}$ is any block-diagonal matrix. We claim the following holds

$$\|y\|_\tau^2 \le \frac{1}{4\alpha}\big\|\left(\|\mathbf{S}_i\|_{\mathrm{op}}\right)_{i\in[m]}\big\|_\tau^2. \tag{37}$$

We postpone its proof to the end. Now we insert $\mathbf{S} = \mathbb{E}[\mathrm{sym}(\mathbf{C}^{-1}\boldsymbol{\Delta}_C)]$ and $\alpha = 2/p - 1$ to the above inequality. This gives the bound for τ-norm

$$\big\|\mathbf{T}^{-1}\mathbf{J}_{\mathbf{C}}[\mathbb{E}\left[\boldsymbol{\Delta}_C\right]]\big\|_\tau \le \sqrt{\frac{p}{2-p}}\cdot\big\|\left(\|\mathbb{E}[\mathrm{sym}\left(\mathbf{C}_i^{-1}\boldsymbol{\Delta}_{C,i}\right)]\|_{\mathrm{op}}\right)_{i\in[m]}\big\|_\tau.$$

For the ℓ_∞-norm, we use the decomposition $\mathbf{R} = \mathbf{D} + \mathbf{K}$ in Lemma 62. We apply the blockwise trace bounds in the proof of Lemma 63 and obtain

$$\begin{aligned}\big\|\mathbf{T}^{-1}\mathbf{J}_{\mathbf{C}}[\mathbb{E}\left[\boldsymbol{\Delta}_C\right]]\big\|_\infty &\le 2\big\|\mathbf{D}\mathbf{T}^{-1}\boldsymbol{\Sigma}\big(\big\|\mathbb{E}\left[\mathrm{sym}\left(\mathbf{C}_i^{-1}\boldsymbol{\Delta}_{C,i}\right)\right]\big\|_{\mathrm{op}}\big)_{i\in[m]}\big\|_\infty \\ &\quad + 2\big\|\mathbf{K}\mathbb{E}\left[l\right]\big\|_\infty + 2\big\|\mathbf{R}\mathbf{T}^{-1}(\mathbf{P}\circ_{\mathrm{G}}\mathbf{P})\mathbb{E}\left[l\right]\big\|_\infty\end{aligned}$$

$$\leq p\left\|\left(\left\|\mathbb{E}\left[\mathrm{sym}\left(\mathbf{C}_i^{-1}\boldsymbol{\Delta}_{C,i}\right)\right]\right\|_{\mathrm{op}}\right)_{i\in[m]}\right\|_{\infty}+O\left(\left\|\mathbb{E}\left[l\right]\right\|_{\mathbf{P}\circ_{\mathrm{G}}\mathbf{P}}\right)$$
$$\leq p\left\|\left(\left\|\mathbb{E}\left[\mathrm{sym}\left(\mathbf{C}_i^{-1}\boldsymbol{\Delta}_{C,i}\right)\right]\right\|_{\mathrm{op}}\right)_{i\in[m]}\right\|_{\infty}+O\left(\beta/C_{\mathrm{norm}}\right).$$

where we use Lemma 28, $0\preceq\mathbf{P}\circ_{\mathrm{G}}\mathbf{P}\preceq\mathbf{T}$ from Lemma 55, and Lemma 62. Combining the above computation, we obtain

$$\begin{aligned}\left\|\mathbf{T}^{-1}\mathbb{E}\left[\delta_\tau\right]\right\|_{\tau+\infty}&\leq\sqrt{p/(2-p)}\left\|\left(\left\|\mathbb{E}\left[\mathrm{sym}\left(\mathbf{C}_i^{-1}\boldsymbol{\Delta}_{C,i}\right)\right]\right\|_{\mathrm{op}}\right)_{i\in[m]}\right\|_{\tau+\infty}+O\left(\beta/C_{\mathrm{norm}}+\beta^2\right)\\&\leq\sqrt{p/(2-p)}\cdot\beta+O\left(C_{\mathrm{norm}}^{-1}+\alpha+\beta\right)\beta.\end{aligned}$$

using Lemma 28 and Lemma 29.

We now prove (37). From the definition of y, it holds that $(\mathbf{T}+\alpha\boldsymbol{\Lambda})y=\mathrm{tr}_{\mathrm{G}}\left(\mathbf{SP}-\mathbf{PSP}\right)$. We compute the inner product with y on both sides and obtain

$$\|y\|_\tau^2=y^\top\mathrm{tr}_{\mathrm{G}}(\mathbf{SP}-\mathbf{PSP})-\alpha\|y\|_{\boldsymbol{\Lambda}}^2.$$

We now expand the first term in the above RHS using Cauchy-Schwarz

$$\begin{aligned}y^\top\mathrm{tr}_{\mathrm{G}}(\mathbf{SP}-\mathbf{PSP})&=\mathrm{tr}\left(\mathsf{diag}_{\mathrm{G}}(y)(\mathbf{SP}-\mathbf{PSP})\right)\\&=\langle(\mathbf{I}-\mathbf{P})\mathsf{diag}_{\mathrm{G}}(y),(\mathbf{I}-\mathbf{P})\mathbf{SP}\rangle\leq\|(\mathbf{I}-\mathbf{P})\mathsf{diag}_{\mathrm{G}}(y)\mathbf{P}\|_{\mathrm{F}}\|(\mathbf{I}-\mathbf{P})\mathbf{SP}\|_{\mathrm{F}}.\end{aligned}$$

In particular, it holds that

$$\begin{aligned}\|(\mathbf{I}-\mathbf{P})\mathsf{diag}_{\mathrm{G}}(y)\mathbf{P}\|_{\mathrm{F}}^2&=\mathrm{tr}(\mathbf{P}\mathsf{diag}_{\mathrm{G}}(y)^2\mathbf{P})-\mathrm{tr}(\mathbf{P}\mathsf{diag}_{\mathrm{G}}(y)\mathbf{P}\mathsf{diag}_{\mathrm{G}}(y)\mathbf{P})\\&=\sum_i\sigma_iy_i^2-\sum_{i,j}y_iy_j\|\mathbf{P}_{ij}\|_{\mathrm{F}}^2=y^\top\boldsymbol{\Lambda}y.\end{aligned}$$

due to $\mathbf{P}^2=\mathbf{P}$ and the definition of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}$. With the similar arguments we compute

$$\|(\mathbf{I}-\mathbf{P})\mathbf{SP}\|_{\mathrm{F}}^2=\mathrm{tr}(\mathbf{PS}(\mathbf{I}-\mathbf{P})\mathbf{SP})\leq\mathrm{tr}(\mathbf{PS}^2\mathbf{P})=\sum_i\mathrm{tr}(\mathbf{S}_i^2\mathbf{P}_{ii})\leq\sum_i\sigma_i\|\mathbf{S}_i\|_{\mathrm{op}}^2$$

Putting everything together, it holds that

$$\|y\|_\tau^2=\|y\|_{\boldsymbol{\Lambda}}\left(\mathrm{tr}(\mathbf{PS}^2\mathbf{P})\right)^{\frac{1}{2}}-\alpha\|y\|_{\boldsymbol{\Lambda}}^2\leq\frac{1}{4\alpha}\mathrm{tr}(\mathbf{PS}^2\mathbf{P})\leq\frac{1}{4\alpha}\left\|\left(\|\mathbf{S}_i\|\right)_{i\in[m]}\right\|_\tau^2$$

due to $\boldsymbol{\Sigma}\preceq\mathbf{T}$. □

Putting the above pieces together, we finally prove the main stability lemma for Lewis weights.

Proof of Lemma 30. The first two claims are true by integrating the results in Lemma 63. The last one is directly from Lemma 65. □

B.3 Stability of μ

In this part, we present the proof of Lemma 31. It requires the following result.

Lemma 66. *Let $(\overline{x},\overline{\tau})$ be an α-approximation of (x,τ). Then the following inequality holds for δ_x,δ_s, g, and other quantities in Algorithm 1*

$$\left\|\nabla^2\Phi(x)^{-\frac{1}{2}}\mathbb{E}\left[\rho^{-1}\mathbf{T}^{-1}\delta_s+\nabla^2\Phi(x)\delta_x\right]-g\right\|_{2,\tau+\infty}\lesssim\beta^2.$$

Proof. We first decompose the i-th block of the target vector

$$\begin{aligned}
&\left(\nabla^2\Phi(x)^{-\frac{1}{2}}\mathbb{E}\left[\rho^{-1}\mathbf{T}^{-1}\delta_s+\nabla^2\Phi(x)\delta_x\right]-g\right)_i=\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\mathbb{E}\left[\delta_{s,i}/(\rho\tau_i)+\nabla^2\phi_i(x_i)\delta_{x,i}\right]-g_i\\
=&\big(\nabla^2\phi_i(x_i)^{\frac{1}{2}}-\nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}}\big)\mathbb{E}[\delta_{x,i}]+\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\big(\nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}}-\nabla^2\phi_i(x_i)^{\frac{1}{2}}\big)\delta_{1,i}\\
&+\tau_i^{-1}(\overline{\tau}_i-\tau_i)\delta_{1,i}+\tau_i^{-1}(\overline{\tau}_i-\tau_i)\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\big(\nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}}-\nabla^2\phi_i(x_i)^{\frac{1}{2}}\big)\delta_{1,i}-\delta_{2,i}
\end{aligned}\tag{38}$$

It requires to bound the ℓ_2 norm of the terms on the RHS except for $\delta_{2,i}$. For the first term in (38), it holds that

$$\left\|\big(\nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}}-\nabla^2\phi_i(x_i)^{\frac{1}{2}}\big)\mathbb{E}[\delta_{x,i}]\right\|_2\leq\alpha\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\mathbb{E}[\delta_{x,i}]\right\|_2$$

due to the fact $\overline{x}$ α-approximates x and Point 2 in Lemma 3. The following two terms involve δ_1 and have the following upper bounds

$$\left\|\tau_i^{-1}(\overline{\tau}_i-\tau_i)\delta_{1,i}\right\|_2=\left|\tau_i^{-1}(\overline{\tau}_i-\tau_i)\right|\left\|\delta_{1,i}\right\|_2\leq\alpha\left\|\delta_{1,i}\right\|_2$$

because the pair $(\overline{x},\overline{\tau})$ is α-approximation of (x,τ) Moreover, it holds that

$$\left\|\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\big(\nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}}-\nabla^2\phi_i(x_i)^{\frac{1}{2}}\big)\delta_{1,i}\right\|_2\leq\left\|\delta_{1,i}\right\|_2\left\|(\overline{x}_i-x_i)\right\|_{\nabla^2\phi_i(x_i)}\leq\alpha\sqrt{n_i}\left\|\delta_{1,i}\right\|_2$$

The additional factor n_i comes from Lemma 22. Due to the same reason, we bound the mixed term

$$\left\|\tau_i^{-1}(\overline{\tau}_i-\tau_i)\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\big(\nabla^2\phi_i(\overline{x}_i)^{\frac{1}{2}}-\nabla^2\phi_i(x_i)^{\frac{1}{2}}\big)\delta_{1,i}\right\|_2\lesssim\alpha\left\|\delta_{1,i}\right\|_2.$$

Combining the above bounds, we have

$$\begin{aligned}
&\left\|\nabla^2\Phi(x)^{-\frac{1}{2}}\mathbb{E}\left[\mathbf{T}^{-1}\delta_s/\rho+\nabla^2\Phi(x)\delta_x\right]-g\right\|_{2,\tau+\infty}\\
&\qquad\lesssim\left\|\delta_2\right\|_{2,\tau+\infty}+\alpha\left\|\delta_1\right\|_{2,\tau+\infty}+\alpha\left\|\nabla^2\Phi(x)^{\frac{1}{2}}\mathbb{E}[\delta_x]\right\|_{2,\tau+\infty}\lesssim\beta^2
\end{aligned}$$

by inserting Lemma 26 and $n_i\alpha\leq\beta$. □

Proof of Lemma 31. We introduce $\kappa_i=(\rho\tau_i)/\rho^{(\mathrm{new})}\tau_i^{(\mathrm{new})}-1$, $\mathrm{SC}_i=\nabla\phi_i(x_i^{(\mathrm{new})})-\nabla\phi_i(x_i)-\nabla^2\phi_i(x_i)\delta_{x_i}$ for the notational convenience. The term $s_i^{(\mathrm{new})}$ is alternatively formalized with the notations as

$$\frac{s_i^{(\mathrm{new})}}{\rho\tau_i}=\frac{s_i}{\rho\tau_i}+\frac{\delta_{s,i}}{\rho\tau_i}=\frac{\delta_{s,i}}{\rho\tau_i}+\mu_i-\nabla\phi_i(x_i).$$

We expand the blockwise difference between $\mu^{(\mathrm{new})}$ and μ

$$\begin{aligned}
\mu_i^{(\mathrm{new})}-\mu_i&=\frac{s_i^{(\mathrm{new})}-s_i}{\rho\tau_i}+\left(\frac{s_i^{(\mathrm{new})}}{\rho^{(\mathrm{new})}\tau_i^{(\mathrm{new})}}-\frac{s_i^{(\mathrm{new})}}{\rho\tau_i}\right)+\nabla^2\phi_i(x_i)\delta_{x,i}+\mathrm{SC}_i\\
&=\left(\frac{\delta_{s,i}}{\rho\tau_i}+\nabla^2\phi_i(x_i)\delta_{x,i}\right)+\mathrm{SC}_i+\kappa_i\left(\frac{\delta_{s,i}}{\rho\tau_i}+\mu_i-\nabla\phi_i(x_i)\right).
\end{aligned}$$

Rearranging and multiplying with $\nabla^2\Phi(x)^{-\frac{1}{2}}$, it holds that

$$\left(\nabla^2\Phi(x)^{-\frac{1}{2}}\delta_\mu-g\right)_i=\left(\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\left(\frac{\delta_{s,i}}{\rho\tau_i}+\nabla^2\phi_i(x_i)\delta_{x,i}\right)-g_i\right)+\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\mathrm{SC}_i\tag{39}$$

$$+\kappa_i \nabla^2\phi_i(x_i)^{-\frac{1}{2}}\left(\frac{\delta_{s,i}}{\rho\tau_i}+\mu_i-\nabla\phi_i(x_i)\right) \tag{40}$$

In order to obtain the claimed bounds, it requires to bound the terms on the above RHS after applying the expectation. We postpone the discussion on the common factor $(1-\kappa_i)^{-1}$. We use the following lemma to bound norm of corresponding full vector for the first parenthesis in (39).

For the first term in (39), we use Lemma 66 to bound the full vector mixed norm

$$\left\|\nabla^2\Phi(x)^{-\frac{1}{2}}\mathbb{E}\left[\rho^{-1}\mathbf{T}^{-1}\delta_s+\nabla^2\Phi(x)\delta_x\right]-g\right\|_{2,\tau+\infty}\lesssim\beta^2.$$

For the second term in (39), it holds that:

$$\begin{aligned}\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\mathbb{E}[\mathrm{SC}_i]\right\|_2=&\left\|\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\mathbb{E}[\nabla\phi_i(x_i^{(\mathrm{new})})-\nabla\phi_i(x_i)-\nabla^2\phi_i(x_i)\delta_{x,i}]\right\|_2\\ \leq&\mathbb{E}\left[\left\|\nabla^2\phi_i(x_i)^{-\frac{1}{2}}[\nabla\phi_i(x_i^{(\mathrm{new})})-\nabla\phi_i(x_i)-\nabla^2\phi_i(x_i)\delta_{x,i}]\right\|_2\right]\\ \leq&\mathbb{E}\left[\frac{\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\right\|_2^2}{(1-\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\right\|_2)^2}\right]\leq 2\mathbb{E}\left[\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\right\|_2^2\right]\end{aligned}$$

where the last line is due to Point 4 Lemma 3 and Point 2 in Lemma 26 We obtain immediately the following full-vector bound

$$\left\|\nabla^2\Phi(x)^{-\frac{1}{2}}\mathbb{E}[\mathrm{SC}]\right\|_{2,\tau+\infty}\lesssim\left\|\left(\mathbb{E}\left[\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\|_2^2\right]\right)_{i\in[m]}\right\|_{\tau+\infty}\lesssim\beta^2$$

using Point 4 in Lemma 26. Now we turn to (40) and first discuss the terms inside the parenthesis. In particular, it holds that $\|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}}=\gamma_i\leq 2\Gamma_i\leq 2\epsilon^2$ as a consequence of Lemma 8. We now compute the blockwise norm

$$\begin{aligned}\left\|\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\left(\mu_i-\nabla\phi_i(x_i)+\frac{\delta_{s,i}}{\rho\tau_i}\right)\right\|_2\leq&\left\|\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\nabla\phi_i(x_i)\right\|_2\\ &+\left\|\mu_i\right\|_{\nabla^2\phi_i(x_i)^{-1}}+\frac{1}{\rho\tau_i}\left\|\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\delta_{s,i}\right\|_2\leq 1+O(C_{\mathrm{norm}}^{-1}+\beta)\end{aligned}$$

where we use the self-concordance, $\epsilon=\theta/C=C_{\mathrm{norm}}^{-1}$, Corollary 27. For κ_i, the following identity holds

$$\kappa_i=\frac{\rho\tau_i-\rho\tau_i^{(\mathrm{new})}}{\rho^{(\mathrm{new})}\tau_i^{(\mathrm{new})}}+\frac{\rho\tau_i^{(\mathrm{new})}-\rho^{(\mathrm{new})}\tau_i^{(\mathrm{new})}}{\rho^{(\mathrm{new})}\tau_i^{(\mathrm{new})}}=\frac{r}{1-r}-\frac{1}{1-r}\frac{\tau_i}{\tau_i^{(\mathrm{new})}}\frac{\delta_{\tau,i}}{\tau_i}.$$

using $\rho^{(\mathrm{new})}=(1-r)\rho$. In particular, the absolute value of the first term admits the following bound

$$\left|\frac{\tau_i}{\tau_i^{(\mathrm{new})}}\frac{\delta_{\tau,i}}{\tau_i}-\frac{\delta_{\tau,i}}{\tau_i}\right|=\frac{\tau_i}{\tau_i^{(\mathrm{new})}}\left(\frac{\delta_{\tau,i}}{\tau_i}\right)^2\leq(1+O(\beta))\left(\frac{\delta_{\tau,i}}{\tau_i}\right)^2$$

where we also use Lemma 26. Therefore, we bound the vector $\kappa=(\kappa_i)_{i\in[m]}$ under the mixed norm, as

$$\|\mathbb{E}[\kappa]\|_{\tau+\infty}\leq(1+O(r))\left\|r\mathbf{1}_m\right\|_{\tau+\infty}+(1+O(r))\left\|\mathbf{T}^{-1}\mathbb{E}[\delta_\tau]\right\|_{\tau+\infty}+O\left(\left\|\mathbb{E}[(\mathbf{T}^{-1}\delta_\tau)^2]\right\|_{\tau+\infty}\right)$$

$$\leq (1+O(r))r(C_{\text{norm}}\sqrt{d}+1)+(1+O(r+\beta))\left(\sqrt{p/(2-p)}\beta+O((\alpha+\beta+C_{\text{norm}}^{-1})\beta)\right)$$
$$\leq 2C_{\text{norm}}r\sqrt{d}+\sqrt{p/(2-p)}\beta+O((\alpha+\beta+C_{\text{norm}}^{-1})\beta)$$

due to Lemma 30 and $r = \alpha\beta/(4C_{\text{norm}}\sqrt{d})$. Since all the terms in the parenthesis of (40) are deterministic, we decompose and estimate

$$\left\|\mathbb{E}\left[\mathsf{diag}_{\mathrm{G}}(\kappa)\nabla^2\Phi(x)^{-\frac{1}{2}}\left(\mu-\nabla\Phi(x)+\rho^{-1}\mathbf{T}^{-1}\delta_s\right)\right]\right\|_{2,\tau+\infty}\leq\left(1+O(C_{\text{norm}}^{-1}+\beta)\right)\|\mathbb{E}[\kappa]\|_{\tau+\infty}$$
$$\leq\sqrt{\frac{p}{2-p}}\beta+O((\alpha+\beta+C_{\text{norm}}^{-1})\beta)+4C_{\text{norm}}r\sqrt{d}\leq\sqrt{\frac{p}{2-p}}\beta+O((\alpha+\beta+C_{\text{norm}}^{-1})\beta).$$

where all the small terms can be absorbed into the second term. Putting the above computations together gives the first claim.

For the second claim, we use the following identity

$$\delta_{\mu,i}=(1+\kappa_i)\frac{\delta_{s,i}}{\rho\tau_i}+\kappa_i\left(\mu_i-\nabla\phi_i(x_i)\right)+\nabla\phi_i(x_i^{(\text{new})})-\nabla\phi_i(x_i). \tag{41}$$

By Lemma 30 and the choice of r, we have $\|\kappa\|_\infty\lesssim\beta$ and $\mathbb{E}[\kappa_i^2]\lesssim r^2+\mathbb{E}[(\tau_i^{-1}\delta_{\tau,i})^2]$. We bound the squared local norms of the three terms separately. For the first term,

$$\left\|\left(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{-\frac{1}{2}}(1+\kappa_i)\rho^{-1}\tau_i^{-1}\delta_{s,i}\|_2^2]\right)_{i\in[m]}\right\|_{\tau+\infty}\lesssim\beta^2,$$

due to Corollary 27, and the α-approximation of $(\overline{x},\overline{\tau})$. For the second term, since $\|\mu_i\|^2_{\nabla^2\phi_i(x_i)^{-1}}=\gamma_i\lesssim\epsilon^2$, we obtain

$$\left\|\left(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\kappa_i\left(\mu_i-\nabla\phi_i(x_i)\right)\|_2^2]\right)_{i\in[m]}\right\|_{\tau+\infty}\lesssim(1+\|\gamma\|_\infty)\left\|\mathbb{E}[\kappa^2]\right\|_{\tau+\infty}$$
$$\lesssim r^2\left\|\mathbf{1}_m\right\|_{\tau+\infty}+\left\|\mathbb{E}[(\mathbf{T}^{-1}\delta_\tau)^2]\right\|_{\tau+\infty}\lesssim\beta^2,$$

where we use the self-concordance, the choice of r, and Lemma 30. For the final term, it holds that

$$\left\|\left(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\left(\nabla\phi_i(x_i^{(\text{new})})-\nabla\phi_i(x_i)\right)\|_2^2]\right)_{i\in[m]}\right\|_{\tau+\infty}\lesssim\left\|\left(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}\delta_{x,i}\|_2^2]\right)_{i\in[m]}\right\|_{\tau+\infty}\lesssim\beta^2,$$

by self-concordance and Point 4 in Lemma 26. Combining these bounds yields

$$\left\|\left(\mathbb{E}[\|\nabla^2\phi_i(x_i)^{-\frac{1}{2}}\delta_{\mu,i}\|_2^2]\right)_{i\in[m]}\right\|_{\tau+\infty}\lesssim\beta^2.$$

Similarly, for the third claim, (41) gives

$$\left\|\nabla^2\Phi(x)^{-\frac{1}{2}}\delta_\mu\right\|_{2,\infty}\lesssim\|\delta_1\|_{2,\infty}+\left(1+\|\gamma\|_\infty^{\frac{1}{2}}\right)\left(r+\left\|\mathbf{T}^{-1}\delta_\tau\right\|_\infty\right)+\left\|\nabla^2\Phi(x)^{\frac{1}{2}}\delta_x\right\|_{2,\infty}\lesssim\beta,$$

by Lemmas 26 and 30.

□

B.4 Proof of Approximate Feasibility

Proof of Lemma 37. We recall from Algorithm 1 that the difference between $x^{(\mathrm{new})}$ and x is expanded as

$$x^{(\mathrm{new})} - x = \nabla^2\Phi(\overline{x})^{-\frac{1}{2}}(g - \mathbf{R}\delta_1 - \mathbf{R}\delta_2)$$

where δ_1 and δ_2 are defined as

$$\begin{aligned}\delta_1 &= \overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g,\\ \delta_2 &= \overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}(\mathbf{A}^\top x - b).\end{aligned}$$

Let us define auxiliary vectors

$$\begin{aligned}\delta_1^\dagger &= \overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\mathbf{H}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g,\\ \delta_2^\dagger &= \overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\mathbf{H}^{-1}(\mathbf{A}^\top x - b).\end{aligned}$$

We use $\delta_1^\dagger$ and $\delta_2^\dagger$ to form a reference vector

$$x^\dagger - x = \nabla^2\Phi(\overline{x})^{-\frac{1}{2}}(g - \delta_1^\dagger - \delta_2^\dagger).$$

From the above definitions, it holds that

$$\begin{aligned}\mathbf{A}^\top x^\dagger &= \mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\left(\mathbf{I} - \overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\mathbf{H}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\right)g\\ &\quad - \mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-1}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\mathbf{H}^{-1}(\mathbf{A}^\top x - b) + \mathbf{A}^\top x\\ &= \mathbf{A}^\top x + \mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}(\mathbf{I} - \mathbf{I})\,g - (\mathbf{A}^\top x - b)\\ &= b.\end{aligned}$$

This suggests that $x^\dagger$ is strictly primal feasible. Then we compute the difference

$$\begin{aligned}\mathbf{A}^\top x^\dagger - \mathbf{A}^\top x^{(\mathrm{new})} &= \mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-1}\left(\mathbf{R}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1} - \nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\mathbf{H}^{-1}\right)\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g\\ &\quad + \mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-1}\left(\mathbf{R}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1} - \nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\mathbf{H}^{-1}\right)(\mathbf{A}^\top x - b)\\ &= \mathbf{H}\left(\overline{\mathbf{H}}^{-1} - \mathbf{H}^{-1}\right)\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g + \mathbf{H}\left(\overline{\mathbf{H}}^{-1} - \mathbf{H}^{-1}\right)(\mathbf{A}^\top x - b)\\ &\quad + \mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-1}(\mathbf{R} - \mathbf{I})\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g\\ &\quad + \mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-1}(\mathbf{R} - \mathbf{I})\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}(\mathbf{A}^\top x - b).\end{aligned}$$

We will compute the $\mathbf{H}^{-1}$-weighted norm of all the terms on the above RHS. For the first one, it holds that

$$\begin{aligned}\left\|\mathbf{H}\left(\overline{\mathbf{H}}^{-1} - \mathbf{H}^{-1}\right)\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g\right\|_{\mathbf{H}^{-1}} &= \left\|\mathbf{H}^{\frac{1}{2}}\left(\overline{\mathbf{H}}^{-1} - \mathbf{H}^{-1}\right)\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g\right\|_2\\ &\le \left\|\mathbf{H}^{\frac{1}{2}}\left(\overline{\mathbf{H}}^{-1} - \mathbf{H}^{-1}\right)\mathbf{H}^{\frac{1}{2}}\right\|_{\mathrm{op}}\left\|\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g\right\|_{\mathbf{H}^{-1}}\\ &\lesssim \alpha\left\|\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g\right\|_{\mathbf{H}^{-1}} \le \alpha\left\|g\right\|_{2,\overline{\tau}} \le \alpha\beta/C_{\mathrm{norm}}\end{aligned}$$

where the second inequality is due to Lemma 4.32 in [vdBLN+20], $\mathbf{H} \approx_\alpha \overline{\mathbf{H}}$, α-approximation, and Lemma 25, the final inequality is due to the norm bound $\|g\|_{2,\tau} \leq C_{\text{norm}}^{-1}\|g\|_{2,\tau+\infty} = \beta$ in Lemma 26, α-approximation, Corollary 25, and approximate feasibility. Similarly, for the second term, it simply holds that

$$\left\|\mathbf{H}\left(\overline{\mathbf{H}}^{-1} - \mathbf{H}^{-1}\right)(\mathbf{A}^\top x - b)\right\|_{\mathbf{H}^{-1}} \lesssim \alpha\left\|\mathbf{A}^\top x - b\right\|_{\mathbf{H}^{-1}} \leq \alpha^2\beta/C_{\text{norm}}.$$

For the third term, it holds that

$$\begin{aligned}
&\left\|\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-1}\left(\mathbf{R}-\mathbf{I}\right)\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g\right\|_{\mathbf{H}^{-1}}\\
=&\left\|\mathbf{H}^{-\frac{1}{2}}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-1}\left(\mathbf{R}-\mathbf{I}\right)\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\mathbf{H}^{-\frac{1}{2}}\mathbf{H}^{\frac{1}{2}}\overline{\mathbf{H}}^{-1}\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g\right\|_2\\
\leq&\left\|\mathbf{H}^{-\frac{1}{2}}\left(\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-1}\mathbf{R}\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}-\mathbf{H}\right)\mathbf{H}^{-\frac{1}{2}}\right\|_{\text{op}}\left\|\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g\right\|_{\overline{\mathbf{H}}^{-1}\mathbf{H}\overline{\mathbf{H}}^{-1}}\\
\leq&3\beta\left\|\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}g\right\|_{\mathbf{H}^{-1}} \lesssim \alpha\|g\|_{2,\overline{\tau}} \leq \alpha\beta/C_{\text{norm}}
\end{aligned}$$

due to Lemma 4.32 in [vdBLN+20], norm bound of g, $\mathbf{H} \approx_\alpha \overline{\mathbf{H}}$, and $\overline{\tau} \approx_\alpha \tau$ again. In a similar fashion, we obtain

$$\left\|\mathbf{A}^\top\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\overline{\mathbf{T}}^{-1}\left(\mathbf{R}-\mathbf{I}\right)\nabla^2\Phi(\overline{x})^{-\frac{1}{2}}\mathbf{A}\overline{\mathbf{H}}^{-1}(\mathbf{A}^\top x - b)\right\|_{\mathbf{H}^{-1}} \lesssim \alpha\left\|\mathbf{A}^\top x - b\right\|_{\mathbf{H}^{-1}} \leq \alpha^2\beta/C_{\text{norm}}.$$

As a consequence of the above computation, it follows that

$$\left\|\mathbf{A}^\top x^{(\text{new})} - b\right\|_{\mathbf{H}^{-1}} \leq \left\|\mathbf{A}^\top x^{(\text{new})} - \mathbf{A}^\top x^\dagger\right\|_{\mathbf{H}^{-1}} + \left\|\mathbf{A}^\top x^\dagger - b\right\|_{\mathbf{H}^{-1}} \lesssim \frac{\alpha\beta}{2C_{\text{norm}}}$$

by our choice of parameters. We choose a sufficiently small constant hidden inside $O(\alpha)$ ratio in the spectral approximation to ensure this. □

B.5 Sampling Scheme Construction

Proof of Lemma 38. The Block-diagonal and Expectation property automatically hold by construction. Also, since R_i and R_j are negatively correlated for any $i \neq j$, $\mathbb{E}[R_iR_j] \leq 1$ holds. For the Variance condition, because every $X^{(j)}$ is i.i.d. sampled, we get

$$\text{Var}(R_i) = \frac{1}{N^2}\sum_{j=1}^{N}\text{Var}\left(X_i^{(j)}\right) \leq \frac{1}{N}\left(\mathbb{E}[X_i^2] - 1\right) \leq \frac{1}{Np_i}.$$

The following holds from the definition of p_i

$$p_i \geq \frac{m}{K\sqrt{d}}\left(\frac{\|\delta_i\|_2^2}{\|\delta\|_2^2} + \frac{1}{m}\right) \geq \frac{1}{K}\sqrt{\frac{m}{d}}\frac{\|\delta_i\|_2}{\|\delta\|_2}. \tag{42}$$

As a consequence, it holds that

$$\text{Var}(R_i\|\delta_i\|_2) = \|\delta_i\|_2^2 \cdot \text{Var}(R_i) \leq \|\delta_i\|_2^2 \cdot \frac{K}{N}\sqrt{\frac{d}{m}}\frac{\|\delta\|_2}{\|\delta_i\|_2} \leq \beta^2\sqrt{\frac{d}{m}}\frac{\|\delta_i\|_2\|\delta\|_2}{C_{\text{Var}}^2}.$$

Also, it holds that $\mathrm{Var}[R_i] \le \tau_i^{-1}$ as $p_i \ge \tau_i/K$. For the Maximum property, we apply Bernstein inequality to $R_i = \frac{1}{N}\sum_{j=1}^{N} X_i^{(j)}$ with $t = \sqrt{\frac{d}{m}}\frac{\beta}{C_{\mathrm{Var}}^2}\frac{\|\delta\|_2}{\|\delta_i\|_2}$

$$\Pr\left[|R_i - 1| \ge t\right] \le 2\exp\left(-\frac{Nt^2/2}{\frac{1-p_i}{p_i} + \frac{t}{3p_i}}\right) \le 2\exp\left(-\frac{1}{4}\min\{Nt^2 p_i, Ntp_i\}\right).$$

Due to $d \ge 1$ and $p_i \ge \frac{m}{K\sqrt{d}}\frac{\|\delta_i\|_2^2}{\|\delta\|_2^2}$, it holds that

$$Nt^2 p_i \ge N\frac{\beta^2\sqrt{d}}{KC_{\mathrm{Var}}^4} \ge 100\sqrt{d}\log n \ge 100\log n.$$

Similarly, because $\beta \le 1$, $C_{\mathrm{Var}} > 1$, and (42), we have

$$Ntp_i \ge N\frac{\beta}{KC_{\mathrm{Var}}^2} \ge 100C_{\mathrm{Var}}^2\beta^{-1}\log n \ge 100\log n.$$

Therefore for any $i \in [m]$, $|R_i - 1| \le \sqrt{\frac{d}{m}}\frac{\beta}{C_{\mathrm{Var}}^2}\frac{\|\delta\|_2}{\|\delta_i\|_2}$ is true w.h.p. As a consequence, w.h.p.

$$\|\mathbf{R}\delta - \delta\|_{2,\infty} \le \max_{i\in[m]}\|\mathbf{R}_i\delta_i - \delta_i\|_2 \le \sqrt{\frac{d}{m}}\frac{\beta}{C_{\mathrm{Var}}^2}\|\delta\|_2$$

Finally, the Spectral Approximation follows from Lemma 2.3 in [LPT+25] and Lemma 61, a trivial adoption of matrix Freedman inequality. □

C Proofs of Data Structures

C.1 Dual Maintenance

Proof of Lemma 40. At each query, FindBlocks will collect all the block indices i's that contain at least one heavy row returned by the block heavy-hitters in Line 12, and combine with indices maintained by "dirty sets" F^l from all dyadic level $l = 0, \ldots, L$. Then VerifyBlocks computes the exact current value of each candidate and resets $\overline{v}_i$ if its error is large. For a fixed dyadic level l, a block index $i \in [m]$ will be added to F_l once its heavy-hitter scaling matrix is updated by the procedure Update, or $\overline{v}_i$ is previously reset at Line 21. Once i is added into F_l, the corresponding scaling in D_l will be set to 0 and this status will remain until the next t with $2^l | t$.

Correctness. Now fix a block index $i \in [m]$. After the t-th call to Query, we want to show

$$\left\|\mathbf{W}_i^{(t)}\left(v_i^{(t)} - \overline{v}_i\right)\right\|_2 \le \epsilon. \tag{43}$$

where $\mathbf{W}_i^{(t)}$ is $\mathbf{W}_i$ just before t-th call. Let t' be the most recent Query call when i is checked. If $t = t'$, (43) trivially holds: either $\overline{v}_i$ is just reset, or $\left\|\mathbf{W}_i^{(t)}\left(v_i^{(t)} - \overline{v}_i\right)\right\|_2 \le 0.2\epsilon$.

If $t' < t$, we decompose $(t', t]$ into contiguous and disjoint dyadic intervals

$$(t_k, t_{k+1}], \qquad t' = t_0 < \cdots < t_K = t$$

Algorithm 6: Dual Maintenance

members
  Block HH D_l in Algorithm 2, vector $\delta^{(l)} \in \mathbb{R}^d$, and set F_l for $l = 0, \ldots, L$
  $\overline{v} \in \mathbb{R}^n$, $\widehat{f} \in \mathbb{R}^d$, block-diagonal $\mathbf{W} \in \mathbb{R}^{n \times n}$
procedure *Initialize*($\mathbf{A} \in \mathbb{R}^{n \times d}, v^{(\text{init})} \in \mathbb{R}^n, \mathbf{W} \in \mathbb{R}^{n \times n}, \epsilon$)
  $\overline{v} \leftarrow v^{(\text{init})}$, $\mathbf{W} \leftarrow \mathbf{W}$, $\widehat{f} \leftarrow 0$, $t \leftarrow 0$
  Execute D_l.Initialize($\mathbf{A}, \mathsf{diag}_{\mathrm{G}}((\sqrt{n_1}, \ldots, \sqrt{n_m}))\mathbf{W}$), and set $\delta^{(l)} \leftarrow 0$, $F_l \leftarrow \emptyset$ for any $l = 0, \ldots, L$
procedure *FindBlocks*(h)
  $I \leftarrow \emptyset$
  for $l = 0, \ldots, L$ **do**
    $\delta^{(l)} \leftarrow \delta^{(l)} + h$
    if $2^l | t$ **then**
      $U \leftarrow D_l$.Query($\delta^{(l)}, 0.2\epsilon / \log T$)
      $I \leftarrow I \cup \{i : \text{Block } i \text{ has at least one coordinate in the set } U\}$
      $I \leftarrow I \cup F_l$
      Execute D_l.ScaleBlock($i, n_i^{1/2} \mathbf{W}_i$) for $i \in F_l$
      $F_l \leftarrow \emptyset$, $\delta^{(l)} \leftarrow 0$
  return I
procedure *VerifyBlocks*(I)
  $I' \leftarrow \emptyset$
  for $i \in I$ **do**
    if $\left\| \mathbf{W}_i \left(v^{(\text{init})} + \mathbf{A}\widehat{f} - \overline{v} \right)_i \right\|_2 \geq 0.2\epsilon$ **then**
      $\overline{v}_i \leftarrow (v^{(\text{init})} + \mathbf{A}\widehat{f})_i$
      Add i into F_l and execute D_l.ScaleBlock($i, 0$) for any $l = 0, \ldots, L$
      $I' \leftarrow I' \cup \{i\}$
  return I'
procedure *Query*(h)
  $t \leftarrow t + 1$
  $\widehat{f} \leftarrow \widehat{f} + h$
  $I \leftarrow$ FindBlocks(h)
  $I' \leftarrow$ VerifyBlocks(I)
  return $\overline{v}, I'$
procedure *Update*($i, \mathbf{V}$)
  Add i into F_l, execute D_l.ScaleBlock($i, 0$) for any $l = 0, \ldots, L$
  $\mathbf{W}_i \leftarrow \mathbf{V}$
procedure *ComputeExact*()

with $K \le 2\log T$ (Lemma 6.3 in [vdBLN+20]). Let interval $(t_k, t_{k+1}]$ corresponds to the dyadic level l_k. Since i is not checked, it is neither added to any F_{l_k} nor detected to be heavy on the interval $(t', t]$. This means $\overline{v}_i$ and $\mathbf{W}_i$ are unchanged, so we can omit the time subscript. Moreover, for any $j \in \mathrm{G}_i$, the row index j is never returned as a heavy row for any $k \in [K]$. By the guarantee of the block heavy-hitters, it holds that

$$\left\| \mathbf{W}_i \mathbf{A}_i \sum_{s=t_k+1}^{t_{k+1}} h^{(s)} \right\|_2 \le \sqrt{n_i} \left\| \mathbf{W}_i \mathbf{A}_i \sum_{s=t_k+1}^{t_{k+1}} h^{(s)} \right\|_\infty \le 0.2\epsilon / \log T.$$

Therefore, the total change in the interval $(t', t]$ is

$$\left\| \mathbf{W}_i \big(v_i^{(t)} - v_i^{(t')} \big) \right\|_2 \le \sum_{k=0}^{K-1} \left\| \mathbf{W}_i \big(v_i^{(t_k)} - v_i^{(t_{k+1})} \big) \right\|_2 \le \frac{0.2K\epsilon}{\log T} \le 0.5\epsilon.$$

As a consequence, the difference between the output $\overline{v}$ and $v^{(t)}$ satisfies

$$\left\| \mathbf{W}_i \big(v_i^{(t)} - \overline{v}_i \big) \right\|_2 \le \left\| \mathbf{W}_i \big(v_i^{(t)} - v_i^{(t')} \big) \right\|_2 + \left\| \mathbf{W}_i \big(v_i^{(t')} - \overline{v}_i \big) \right\|_2 \le 0.5\epsilon + 0.2\epsilon < \epsilon$$

because $\overline{v}_i$ is either reset at t', or is an 0.2ϵ-approximation of $v_i^{(t')}$ under weighted norm.

Runtime. We analyze the total running time of T calls to QUERY. We first bound the time spent in FINDBLOCKS. Fix a dyadic level l, and consider the current interval $(t-2^l, t]$ for some t satisfying $2^l | t$. If $i \in F_l$ at t, then its heavy hitter block scaling has been set to zero. Therefore, the i-th block does not contribute to the movement and the running time of heavy-hitter D_l. Otherwise, if $i \notin F_l$, its heavy-hitter scaling remains $n_i^{\frac{1}{2}} \mathbf{W}_i$ for the entire interval. Define diagonal mask matrix $\mathbf{F} \in \mathbb{R}^{n \times n}$ such that $\mathbf{F}_i = \mathbf{I}_{n_i}$ for $i \notin F_l$ and 0 otherwise. Note that we will charge the cost of restoring heavy-hitter scaling in Line 14 to the update and reset event discussed below. Combining the running time cost of heavy-hitter query with the cost of updating $\widehat{f}$, we bound the amortized time spent by FINDBLOCKS at level l over the time interval $[t-2^l, t]$

$$\widetilde{O}\left(d\epsilon^{-2} \left\| \mathbf{F}\mathbf{W}^{(t)} \mathbf{A} \delta^{(l)} \right\|_2^2 + d \right) \le \widetilde{O}\left(2^l \epsilon^{-2} \sum_{s=t-2^l+1}^{t} \left\| \mathbf{W}^{(s)} \mathbf{A} h^{(s)} \right\|_2^2 + 2^l d \right),$$

The line follows from triangle inequality and Cauchy-Schwarz, also because $\mathbf{W}_i$ remains unchanged throughout the interval for every unmasked block. Similarly, as mapping coordinates to blocks can only decrease the cardinality, the number of block heavy entries added to I in Line 12 is at most $\widetilde{O}\left(\epsilon^{-2} \left\| \mathbf{F}\mathbf{W}\mathbf{A} \delta^{(l)} \right\|_2^2 \right) \le \widetilde{O}\left(2^l d\epsilon^{-2} \sum_{s=t-2^l+1}^{t} \left\| \mathbf{W}^{(s)} \mathbf{A} h^{(s)} \right\|_2^2 \right)$. We will use this fact later. The running time for T calls and all l is

$$\sum_{l=0}^{L} \sum_{t \in [T]: 2^l | t} \widetilde{O}\left(2^l d\epsilon^{-2} \sum_{s=t-2^l+1}^{t} \left\| \mathbf{W}^{(s)} \mathbf{A} h^{(s)} \right\|_2^2 + 2^l d \right) \le \widetilde{O}\left(Td + Td\epsilon^{-2} \sum_{t \in [T]} \left\| \mathbf{W}^{(t)} \mathbf{A} h^{(t)} \right\|_2^2 \right). \tag{44}$$

We next bound the time spent by VERIFYBLOCKS. For every single call t and index set I to be checked, it requires $\widetilde{O}(d)$ time to materialize v_i in Line 20 for any $i \in I$. We distinguish

three sources of candidates in I. First, i is added to I by some heavy-hitter D_l. Previous analysis indicates the number of such blocks detected by a single D_l on the interval $(t-2^l, t]$ is at most $\widetilde{O}\left(2^l\epsilon^{-2}\sum_{s=t-2^l+1}^{t}\left\|\mathbf{W}\mathbf{A}h^{(s)}\right\|_2^2\right)$. The total time cost across T calls equals $\widetilde{O}(d)$ times the number of block heavy entries detected over all $l=0,\ldots,L$ and all iterations. Therefore, the contribution due to the first source is also bounded by (44) with the same calculation.

For the second source, i is included through F_l because its heavy-hitter scaling is changed after a call to Update. Each call inserts i into every F_l and causes L future block checks. Since $L \le O(\log T)$, we attribute $\widetilde{O}(d)$ cost to the call to Update by combining the check cost and the cost of rescaling in Line 14.

For the third resource, i is included through F_l because $\overline{v}_i$ is reset at a previous call to Query. Similar to the second, each reset event causes L future block checks. We will charge the $\widetilde{O}(d)$ costs to the reset event. Fix i and let $1 \le t_1 < \cdots < t_k < \ldots t_K \le T$ be reset times of $\overline{v}_i$. Consider an interval $(t_{k-1}, t_k]$. If Update is called for i during the interval, we already attribute the reset cost to Update. Otherwise, $\mathbf{W}_i$ remains unchanged. As a consequence, the total number of resets with no preceding call to Update is controlled as

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{1}[i \text{ reset at } t, \text{ not updated since last reset}] \le & \sum_{k:\text{ no updates}}\mathbb{1}\left[\left\|\mathbf{W}_i^{(t_k)}\mathbf{A}_i\sum_{s=t_{k-1}+1}^{t_k}h^{(s)}\right\|_2 \ge 0.2\epsilon\right] \\
\le & \widetilde{O}\left(\epsilon^{-2}\sum_{k:\text{ no updates}}\left\|\mathbf{W}_i^{(t_k)}\mathbf{A}_i\sum_{s=t_{k-1}+1}^{t_k}h^{(s)}\right\|_2^2\right) \\
\le & \widetilde{O}\left(T\epsilon^{-2}\sum_{t=1}^{T}\left\|\mathbf{W}_i^{(t)}\mathbf{A}_i h^{(t)}\right\|_2^2\right),
\end{aligned}$$

where the last line uses Cauchy-Schwarz. By summing over $i \in [m]$, the amortized time due to the third source is $\widetilde{O}\left(\epsilon^{-2}Td\sum_{t=1}^{T}\left\|\mathbf{W}^{(t)}\mathbf{A}h^{(t)}\right\|_2^2\right)$.

Since $n_i \le O(1)$ for any $i \in [m]$, the runtime of Initialize is $\widetilde{O}(nd+d^2)$. Moreover, each call to Update takes $\widetilde{O}(d)$ amortized time when we combine with the amortized cost of the block checks. Note that we adapt a ComputeExact subroutine from [vdBLSS20] that runs in $\widetilde{O}(n)$ time. □

C.2 Gradient Maintenance

Proof of Lemma 41. For any $i \in [m]$, we decompose $(\nabla_\mu^{\text{loc}}\Psi(\mu,x))_i = a_i u_i$ where

$$a_i = \left\|(\nabla_\mu^{\text{loc}}\Psi(\mu,x))_i\right\|_2, \qquad u_i = (\nabla_\mu^{\text{loc}}\Psi(\mu,x))_i/a_i.$$

The norm-direction decoupling allows the following dual vector presentation.

Claim 67. *For any $x \in \mathbb{R}^n$ with $x_i = a_i u_i$, where $a_i \in \mathbb{R}_{\ge 0}$ is a scalar and $u_i \in \mathbb{R}^{n_i}$ is a unit vector, the dual vector $x^{\flat(\tau)}$ is expressed as*

$$x_i^{\flat(\tau)} = \begin{cases} a_i^{\flat(\tau)}u_i, & a_i \ne 0, \\ 0, & a_i = 0, \end{cases}$$

where $a^{\flat(\tau)}$ is the dual vector of a.

Algorithm 7: Gradient Maintenance

members
  $v^{(k,l)} \in \{0,1\}^m$, $w^{(k,l)} \in \mathbb{R}^d$, $u^{(k,l)} \in \mathbb{R}^n$
  $\mu \in \mathbb{R}^n$, $x \in \mathbb{R}^n$, $\tau \in \mathbb{R}^m$, $\overline{\tau} \in \mathbb{R}^m$
  $a, \overline{a} \in \mathbb{R}^m$, $u \in \mathbb{R}^n$
procedure *Initialize*($\mathbf{A} \in \mathbb{R}^{n \times d}, \mu \in \mathbb{R}^n, x \in \mathbb{R}^n, \tau \in \mathbb{R}^m, \epsilon$)
  $\mathbf{A} \leftarrow \mathbf{A}$, $\mu \leftarrow \mu$, $x \leftarrow x$, $\tau \leftarrow \tau$, $a_{\max} \leftarrow 10\lambda\delta e^{10\lambda\delta^2}$, $a_{\min} \leftarrow \lambda\epsilon$
  for $i \in [m]$ **do**
    $u_i \leftarrow \left(\nabla_\mu^{\mathrm{loc}} \Psi(\mu, x)\right)_i$
    $a_i \leftarrow \|u_i\|_2$
    Find k such that $(1-\epsilon)^k \le \tau_i/(2d) < (1-\epsilon)^{k-1}$, and set $\overline{\tau}_i \leftarrow 2d(1-\epsilon)^k$
    if $a_i \le a_{\min}$ **then**
      $\overline{a}_i \leftarrow 0$, $u_i \leftarrow 0$
    else
      $u_i \leftarrow u_i / a_i$
      Find l such that $(1-\epsilon)^l \le a_i/a_{\max} < (1-\epsilon)^{l-1}$, and set $\overline{a}_i \leftarrow a_{\max}(1-\epsilon)^l$
  for $k = 1, \ldots, O(\epsilon^{-1}\log(m))$, *and* $l = 1, \ldots, O(\epsilon^{-1}\log(a_{\max}/a_{\min}))$ **do**
    $v^{(k,l)} \leftarrow 0$
    $v_i^{(k,l)} \leftarrow 1$ if $\overline{a}_i = a_{\max}(1-\epsilon)^l$ and $\overline{\tau}_i = 2d(1-\epsilon)^k$
    $u^{(k,l)} \leftarrow \sum_{i : v_i^{(k,l)} = 1} u_i$
    $w^{(k,l)} \leftarrow \sum_{i : v_i^{(k,l)} = 1} \mathbf{A}_i^\top \nabla^2 \phi_i(x_i)^{-\frac{1}{2}} u_i$
procedure *Update*($i \in [m], \nu, y, \sigma$)
  if $\overline{a}_i > 0$ **then**
    Find k, l such that $v_i^{(k,l)} = 1$, then set $v_i^{(k,l)} = 0$
    $u^{(k,l)} \leftarrow u^{(k,l)} - u_i$
    $w^{(k,l)} \leftarrow w^{(k,l)} - \mathbf{A}_i^\top \nabla^2 \phi_i(x_i)^{-\frac{1}{2}} u_i$
  $\mu_i \leftarrow \nu$, $x_i \leftarrow y$, $\tau_i \leftarrow \sigma$
  $u_i \leftarrow \left(\nabla_\mu^{\mathrm{loc}} \Psi(\mu, x)\right)_i$
  $a_i \leftarrow \|u_i\|_2$
  Find k such that $(1-\epsilon)^k \le \tau_i/(2d) < (1-\epsilon)^{k-1}$, and set $\overline{\tau}_i \leftarrow 2d(1-\epsilon)^k$
  if $a_i \le a_{\min}$ **then**
    $\overline{a}_i \leftarrow 0$, $u_i \leftarrow 0$
  else
    $u_i \leftarrow u_i / a_i$
    Find l such that $(1-\epsilon)^l \le a_i/a_{\max} < (1-\epsilon)^{l-1}$, and set $\overline{a}_i \leftarrow a_{\max}(1-\epsilon)^l$
    $v_i^{(k,l)} \leftarrow 1$, $u^{(k,l)} \leftarrow u^{(k,l)} + u_i$
    $w^{(k,l)} \leftarrow w^{(k,l)} + \mathbf{A}_i^\top \nabla^2 \phi_i(x_i)^{-\frac{1}{2}} u_i$
procedure *Query*()
  Obtain $s = \operatorname{argmin}_{\|w\|_2 + \left\|w/\sqrt{\overline{\tau}}\right\|_\infty \le 1} \left\langle w, \overline{a}/\sqrt{\overline{\tau}} \right\rangle = \sum_{k,l} s^{(k,l)} v^{(k,l)}$ using Algorithm 8 from [LS19]
  return $\sum_{k,l} s^{(k,l)} u^{(k,l)}$, $\sum_{k,l} s^{(k,l)} w^{(k,l)}$

Moreover, it holds that $a_i = 2\lambda e^{\lambda\Gamma_i}\|\Delta_{\phi_i^*}\|_{\nabla^2\phi_i(x_i)} \le a_{\max} := 10\lambda\delta e^{10\lambda\delta^2}$. Set $a_{\min} := \lambda\epsilon$. For each $a_i > a_{\min}$, choose the smallest k, l such that $\overline{a}_i = a_{\max}(1-\epsilon)^l \le a_i$ and $\overline{\tau}_i = 2d(1-\epsilon)^k \le \tau_i$. If $a_i \le a_{\min}$, we round $\overline{a}_i$ to zero and set $u_i = 0$. This is valid by choosing $\overline{\mu}_i = 0$. Indeed, since $a_i = 2\lambda e^{\lambda\Gamma_i}\|\Delta_{\phi_i^*}\|_{\nabla^2\phi_i(x_i)}$, we have $\|\Delta_{\phi_i^*}\|_{\nabla^2\phi_i(x_i)} \le \epsilon/2$. By self-concordance,

$$\|\overline{\mu}_i - \mu_i\|_{\nabla^2\phi_i(x_i)^{-1}} \le \frac{\|\Delta_{\phi_i^*}\|_{\nabla^2\phi_i(x_i)}}{1 - \|\Delta_{\phi_i^*}\|_{\nabla^2\phi_i(x_i)}} \le \frac{\epsilon}{2-\epsilon} < \epsilon.$$

Moreover, $(\nabla^{\mathrm{loc}}_\mu \Psi(\overline{\mu}, x))_i = 0$. Algorithm 7 then partitions all blocks with $\overline{a}_i > 0$ into disjoint buckets $v^{(k,l)}$ according to the rounding. We then invoke Algorithm 8 in [LS19] to compute the bucket presentation $s^{(k,l)}$ of the dual $\overline{a}^{\flat\overline{\tau}}$, as blocks within the same bucket have the same dual coefficient. Lifting it gives the dual gradient $\sum_{k,l} s^{(k,l)} u^{(k,l)}$, where $u^{(k,l)}$ is the aggregated direction of all u_i in the same bucket.

Correctness. We want to show that there exists some $\overline{\mu}$ such that $\nabla^{\mathrm{loc}}_\mu \Psi(\overline{\mu}, x)$ equals the rounded primal presentation $\sum_{k,l} a_{\max}(1-\epsilon)^l u^{(k,l)}$, and also satisfies $\|\overline{\mu}_i - \mu_i\|_{\nabla^2\phi_i(x_i)^{-1}} \le \epsilon$ for any $i \in [m]$. If this claim is true, by the previous dual presentation claim and the above cited lemma, it holds that

$$(\nabla^{\mathrm{loc}}_\mu \psi(\overline{\mu}, x))^{\flat(\overline{\tau})} = \sum_{kl} s^{(k,l)} u^{(k,l)}, \qquad \mathbf{A}^\top \nabla^2\Phi(x)^{-\frac{1}{2}} (\nabla^{\mathrm{loc}}_\mu \psi(\overline{\mu}, x))^{\flat(\overline{\tau})} = \sum_{kl} s^{(k,l)} w^{(k,l)}.$$

For any $i \in [m]$, it requires to show such a $\overline{\mu}_i$ exists and is close to μ_i in local distance. Let $\Delta^*_{\phi_i} = \nabla\phi_i^*(\nabla\phi(x_i) - \mu_i) - x_i$. Due to primal-dual derivative identities from Lemma 5, μ_i can be expressed through

$$\mu_i = \nabla\phi_i(x_i) - \nabla\phi_i\big(x_i + \Delta^*_{\phi_i}\big).$$

Because ϕ_i is self-concordant and $x_i, x_i + \Delta^*_{\phi_i} \in \operatorname{dom}\phi_i$, the linear interpolation $x_i + t\Delta^*_{\phi_i}$ lies in $\operatorname{dom}\phi_i$. As a consequence, we can legally define the following

$$\mu_i(t) = \nabla\phi_i(x_i) - \nabla\phi_i\big(x_i + t\Delta^*_{\phi_i}\big)$$

for any $t \in [0,1]$. Now let t_i be the solution of the following equation of t:

$$\overline{a}_i/a_i = t\exp\left(\lambda(\Gamma_i(t) - \Gamma_i(1))\right), \qquad \text{where} \quad \Gamma_i(t) = 2D_{\phi_i^*}(\nabla\phi_i(x_i) - \mu_i(t), \nabla\phi_i(x_i)). \tag{45}$$

We claim that $\overline{\mu}_i := \mu(t_i)$ satisfies $\overline{a}_i u_i = \big(\nabla^{\mathrm{loc}}\Psi(\overline{\mu}, x)\big)_i$ and hence the output of QUERY is based on the dual vector $\nabla^{\mathrm{loc}}\Psi(\overline{\mu}, x)^{\flat(\overline{\tau})}$. $\Gamma_i(t)$ is non-decreasing due to

$$\Gamma_i'(t) = 2t\Delta_{\phi_i^*}^\top \nabla^2\phi(x_i + t\Delta_{\phi_i^*})\Delta_{\phi_i^*} \ge 0.$$

As a consequence of the identity in (45) and monotonicity, t_i is in $[\overline{a}_i/a_i, 1]$ and hence can be legally defined. Therefore we compute

$$\begin{aligned}\big(\nabla^{\mathrm{loc}}\Psi(\overline{\mu}, x)\big)_i &= -2\lambda e^{\lambda\Gamma_i(t_i)}\nabla^2\phi(x_i)^{\frac{1}{2}}(\nabla\phi_i^*(\nabla\phi_i(x_i) - \overline{\mu}_i) - x_i) \\ &= -t_i e^{\lambda(\Gamma_i(t_i) - \Gamma_i(1))} \cdot 2\lambda e^{\lambda\Gamma_i(1)}\nabla^2\phi(x_i)^{\frac{1}{2}}\Delta^*_{\phi_i} = \frac{\overline{a}_i}{a_i}\big(\nabla^{\mathrm{loc}}\Psi(\mu, x)\big)_i = \overline{a}_i u_i.\end{aligned}$$

We next show that $\overline{\mu}_i$ is close to μ_i. The time derivative of $\mu_i(t)$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_i(t) = -\nabla^2\phi_i(x_i + t\Delta_{\phi_i^*})\Delta_{\phi_i^*}.$$

As a consequence, we bound the weighted norm as

$$\begin{aligned}\left\|\overline{\mu}_i - \mu_i\right\|_{\nabla^2\phi(x_i)^{-1}} = \left\|\overline{\mu}_i(t_i) - \mu_i(1)\right\|_{\nabla^2\phi(x_i)^{-1}} \leq & \int_{t_i}^1 \left\|\nabla^2\phi_i(x_i + t\Delta_{\phi_i^*})\Delta_{\phi_i^*}\right\|_{\nabla^2\phi(x_i)^{-1}}\mathrm{d}t \\ \leq & \|\Delta_{\phi_i^*}\|_{\nabla^2\phi(x_i)}\int_{t_i}^1 \frac{1}{\left(1 - t\|\Delta_{\phi_i^*}\|_{\nabla^2\phi(x_i)}\right)^2}\mathrm{d}t \\ \leq & 2(1-t_i)\|\Delta_{\phi_i^*}\|_{\nabla^2\phi(x_i)} \leq 4\delta\frac{a_i - \overline{a}_i}{a_i} \leq 4\delta\epsilon\end{aligned}$$

where the second line is due to self-concordance and Lemma 3, and the final line is due to $t_i \in [\overline{a}_i/a_i, 1]$, the gap bound for $(a_i - \overline{a}_i)/a_i \leq \epsilon$, and also $\|\Delta_{\phi_i^*}\|_{\nabla^2\phi(x_i)} \leq 2\left\|\mu_i\right\|_{\nabla^2\phi_i(x_i)^{-1}} \leq 2\delta$. As $\delta \leq 1/10$, we conclude that $\|\mu_i - \overline{\mu}_i\|_{\nabla^2\phi_i(x_i)^{-1}} \leq \epsilon$.

Runtime. Each update will change i from certain bucket $v^{(k,l)}$ to a different one. Updating $w^{(k,l)}$ and $u^{(k,l)}$ correspondingly will take $n_i^2 d + n_i$ time, which is still $O(d)$ when we have $n_i = O(1)$ for any $i \in [m]$. Each query takes $\widetilde{O}(n)$ time majorly due to the call of Algorithm 8 in [LS19]. □

C.3 Leverage Score Maintenance

Proof of Lemma 44. The roles and logic of FindBlocks, VerifyBlocks, and F_l are identical to those in the Dual Maintenance described in Lemma 40, except that $\overline{\sigma}_i$ is compared with and reset to a high-accuracy approximation of the real σ_i. Moreover, we choose sufficiently large constants, sampling rates for $\mathbf{S}_f$ and $\mathbf{S}_v$ ensure the following spectral approximations of the Gram matrices in Line 16 and 27

$$\mathbf{A}^\top\mathbf{V}^2\mathbf{S}_f^2\mathbf{A} \approx_{0.02\epsilon/\log T} \mathbf{A}^\top\mathbf{V}^2\mathbf{A}, \qquad \mathbf{A}^\top\mathbf{V}^2\mathbf{S}_v^2\mathbf{A} \approx_{0.02\epsilon} \mathbf{A}^\top\mathbf{V}^2\mathbf{A}. \tag{46}$$

The first one holds also when we replace $\mathbf{V}$ with $\mathbf{V} - \mathbf{\Delta}^{(l)}$. The column numbers of the JL matrices $\mathbf{R}_f$ and $\mathbf{R}_v$ are chosen properly to ensure their respective approximation rates.

Correctness. Fix block index $i \in [m]$. After the t-th call to Query(), we want to show that

$$\overline{\sigma}_i \approx_\epsilon \sigma(\mathbf{V}^{(t)}\mathbf{A})_i + \overline{z} \tag{47}$$

where $\mathbf{V}^{(t)}$ is $\mathbf{V}$ just before the t-th call. The following holds for any row index $j \in [n]$

$$\|e_j^\top\mathbf{V}^{(t)}\mathbf{A}\mathbf{H}_v\|_2^2 \approx_{0.02\epsilon} \|e_j^\top\mathbf{V}^{(t)}\mathbf{A}\mathbf{M}\mathbf{A}^\top\mathbf{V}^{(t)}\mathbf{S}_v\mathbf{R}_v\|_2^2 \approx_{0.06\epsilon} \|e_j^\top\mathbf{P}(\mathbf{V}^{(t)}\mathbf{A})\|_2^2 \tag{48}$$

due to the inverse accuracy $\mathbf{M} \approx_{0.02\epsilon/\log T} (\mathbf{A}\mathbf{V}^2\mathbf{S}_v^2\mathbf{A})^{-1}$, (46), and also because JL-matrix $\mathbf{R}_v$ ensures an 0.05ϵ-approximation. Let t' be the most recent call to Query when i is checked. We consider the case $t' = t$. If $\overline{\sigma}_i$ is just reset, we obtain the following as a consequence of (48)

$$\left|\sigma(\mathbf{V}^{(t)}\mathbf{A})_i + \overline{z} - \overline{\sigma}_i\right| \leq \sum_{j\in\mathrm{G}_i}\left|\left\|e_j^\top\mathbf{V}^{(t)}\mathbf{A}\mathbf{H}_v\right\|_2^2 - \left\|e_j^\top\mathbf{P}(\mathbf{V}^{(t)}\mathbf{A})\right\|_2^2\right| \leq 0.1\epsilon\sigma(\mathbf{V}^{(t)}\mathbf{A})_i.$$

Algorithm 8: Block Leverage Score Maintenance

members
  $\mathbf{V} \in \mathbb{R}^{n \times n}$, block heavy hitter D_l, matrix $\mathbf{\Delta}^{(l)} \in \mathbb{R}^{n \times n}$, set $F_l \subset [m]$ for $l = 0, \ldots, L$
procedure *Initialize*($\mathbf{A} \in \mathbb{R}^{n \times d}, \mathbf{V} \in \mathbb{R}^{n \times n}, \epsilon$)
  $\mathbf{V} \leftarrow \mathbf{V}$, $t \leftarrow 0$, $L \leftarrow \log T$, $\overline{z} \leftarrow d/m$
  $z_i \leftarrow n_i^{-1} \overline{z}$ for any $i \in [m]$
  for $l = 0, \ldots, L$ **do**
    Execute D_l.Initialize($\mathbf{A}, \mathsf{diag}_{\mathrm{G}}(z^{-\frac{1}{2}})\mathbf{V}$), and set $F_l \leftarrow \emptyset$, $\mathbf{\Delta}^{(l)} \leftarrow 0$ for any l
  $\overline{\sigma} \approx_{0.1\epsilon} \sigma(\mathbf{VA}) + \overline{z}\mathbf{1}_m$
  return $\overline{\sigma}$
procedure *FindBlocks*()
  $I \leftarrow \emptyset$
  for $l = 0, \ldots, L$ **do**
    if $2^l | t$ **then**
      $S_i = 1$ for $i \in F_l$, and set $S_i = 1/\sqrt{p_i}$ with probability $p_i = \min(1, \widetilde{O}(\epsilon^{-2}\overline{\sigma}_i))$, otherwise $S_i = 0$ for all $i \notin F_l$; $\mathbf{S}_f \leftarrow \mathsf{diag}_{\mathrm{G}}(S)$
      $\mathbf{R}_f \in \mathbb{R}^{n \times r_f}$ as a 1/2-approximation JL-matrix
      $\mathbf{M}^{(\mathrm{old})} \approx_{0.02/d} (\mathbf{A}^\top (\mathbf{V} - \mathbf{\Delta}^{(l)})^2 \mathbf{S}_f^2 \mathbf{A})^{-1}$, $\mathbf{M}^{(\mathrm{new})} \approx_{0.02/d} (\mathbf{A}^\top \mathbf{V}^2 \mathbf{S}_f^2 \mathbf{A})^{-1}$
      $\mathbf{H}_f \leftarrow \mathbf{M}^{(\mathrm{new})} \mathbf{A}^\top \mathbf{V} \mathbf{S}_f \mathbf{R}_f - \mathbf{M}^{(\mathrm{old})} \mathbf{A}^\top (\mathbf{V} - \mathbf{\Delta}^{(l)}) \mathbf{S}_f \mathbf{R}_f$
      $U \leftarrow \cup_{k \in [r_f]} D_l$.Query($\mathbf{H}_f e_k, 0.02\epsilon/(r_f \log T)$)
      $I \leftarrow I \cup \{i :$ Block i has at least one coordinate in the set $U\} \cup F_l$
      Execute D_l.ScaleBlock($i, z_i^{-1/2} \mathbf{V}_i$) for $i \in F_l$
      $F_l \leftarrow \emptyset$, $\mathbf{\Delta}^{(l)} \leftarrow 0$,
  return I
procedure *VerifyBlocks*(I)
  $I' \leftarrow \emptyset$
  $S_i = 1/\sqrt{p_i}$ with probability $p_i = \min(1, \widetilde{O}(\epsilon^{-2}\overline{\sigma}_i))$, otherwise $S_i = 0$ for any $i \in [m]$; $\mathbf{S}_v \leftarrow \mathsf{diag}_{\mathrm{G}}(S)$
  $\mathbf{R}_v \in \mathbb{R}^{n \times r_v}$ as a 0.02ϵ-approximate JL-matrix
  $\mathbf{M} \approx_{0.02\epsilon/\log T} (\mathbf{A}^\top \mathbf{V}^2 \mathbf{S}_v^2 \mathbf{A})^{-1}$, and then $\mathbf{H}_v = \mathbf{M} \mathbf{A}^\top \mathbf{V} \mathbf{S}_v \mathbf{R}_v$
  for $i \in I$ **do**
    if $\overline{\sigma}_i \not\approx_{0.3\epsilon} \sum_{j \in \mathrm{G}_i} \|e_j^\top \mathbf{VAH}_v\|_2^2 + \overline{z}$ **then**
      $\overline{\sigma}_i \leftarrow \sum_{j \in \mathrm{G}_i} \|e_j^\top \mathbf{VAH}_v\|_2^2 + \overline{z}$
      Add i into F_l and execute D_l.ScaleBlock($i, 0$) for any l
      $I' \leftarrow I' \cup \{i\}$
  return I'
procedure *Scale*($i, \mathbf{W}$)
  Add i into F_l, execute D_l.ScaleBlock($i, 0$) and $\mathbf{\Delta}_i^{(l)} \leftarrow \mathbf{\Delta}_i^{(l)} + \mathbf{W} - \mathbf{V}_i$ for any l
  $\mathbf{V}_i \leftarrow \mathbf{W}$
procedure *Query()*
  $t \leftarrow t + 1$
  $I \leftarrow$ FindBlocks()
  $I' \leftarrow$ VerifyBlocks(I)
  return $\overline{\sigma}, I'$

This immediately gives (47). Otherwise, $\overline{\sigma}_i$ is already a 0.3ϵ-approximation of $\sum_{j\in \mathrm{G}_i} \left\|e_j^\top \mathbf{V}^{(t)}\mathbf{A}\mathbf{H}_v\right\|_2^2 + \overline{z}$. This also implies (47) by the triangle inequality.

Now we consider the case $t' < t$ and partition $(t', t]$ into contiguous and disjoint dyadic intervals

$$(t_k, t_{k+1}], \qquad t' = t_0 < \cdots < t_K = t$$

with $K \le 2\log T$. Let interval $(t_k, t_{k+1}]$ correspond to the dyadic level l_k. Since i is not checked, it is neither added to any F_l nor detected as heavy on the interval $(t', t]$. This means v_i is unchanged. We now claim that

$$\sigma(\mathbf{V}^{(t_{k+1})}\mathbf{A})_i + \overline{z} \approx_{0.25\epsilon/\log T} \sigma(\mathbf{V}^{(t_k)}\mathbf{A})_i + \overline{z}. \tag{49}$$

holds for any k. We postpone the proof to the end of the part. With (49) being true, it holds that

$$\sigma(\mathbf{V}^{(t')}\mathbf{A})_i + \overline{z} = \sigma(\mathbf{V}^{(t_0)}\mathbf{A})_i + \overline{z} \approx_{0.5\epsilon} \sigma(\mathbf{V}^{(t_K)}\mathbf{A})_i + \overline{z} = \sigma(\mathbf{V}^{(t)}\mathbf{A})_i + \overline{z}$$

as there are at most $K \le 2\log T$ dyadic intervals between t' and t. Consequently, we have

$$\overline{\sigma}_i \approx_{0.3\epsilon} \sum_{j\in \mathrm{G}_i} \left\|e_j^\top \mathbf{V}^{(t')}\mathbf{A}\mathbf{H}_v\right\|_2^2 + \overline{z} \approx_{0.1\epsilon} \sigma(\mathbf{V}^{(t')}\mathbf{A})_i + \overline{z} \approx_{0.2\epsilon} \sigma(\mathbf{V}^{(t)}\mathbf{A})_i + \overline{z}$$

because $\overline{\sigma}_i$ is lastly checked at t', and $\|e_j^\top(\mathbf{V}^{(t')})^\top\mathbf{A}\mathbf{H}_v\|_2^2 \approx_{0.1\epsilon} \|e_j^\top\mathbf{P}(\mathbf{V}^{(t')}\mathbf{A})\|_2^2$ is also true. This immediately implies (47).

We now prove (49). Fix a dyadic interval $(t_{k-1}, t_k]$ corresponding to level l, and consider the t_k-th call. Since block i is not checked, none of its rows $j \in \mathrm{G}_i$ is returned by heavy-hitters. Let $\mathbf{F}$ denote the mask matrix associated with F_l and write $\mathbf{B}^{(\mathrm{new})} = \mathbf{M}^{(\mathrm{new})}\mathbf{A}^\top\mathbf{V}^{(t_k)}$, $\mathbf{B}^{(\mathrm{old})} = \mathbf{M}^{(\mathrm{old})}\mathbf{A}^\top(\mathbf{V}^{(t_k)} - \mathbf{\Delta}^{(l)}) = \mathbf{M}^{(\mathrm{old})}\mathbf{A}^\top\mathbf{V}^{(t_{k-1})}$. By the heavy-hitter guarantees, it holds that

$$\begin{aligned}\left\|e_j^\top\mathbf{F}\mathbf{Z}^{-\frac{1}{2}}\mathbf{V}^{(\mathrm{old})}\mathbf{A}\big(\mathbf{B}^{(\mathrm{new})} - \mathbf{B}^{(\mathrm{old})}\big)\mathbf{S}_f\right\|_2^2 &\le 3\left\|e_j^\top\mathbf{F}\mathbf{Z}^{-\frac{1}{2}}\mathbf{V}^{(\mathrm{old})}\mathbf{A}\mathbf{H}_f\right\|_2^2 \\ &\le 3\sum_{s=1}^{r}\left|e_j^\top\mathbf{F}\mathbf{Z}^{-\frac{1}{2}}\mathbf{V}^{(\mathrm{old})}\mathbf{A}\mathbf{H}_f e_s\right|^2 \le 0.0012\epsilon^2/\log^2 T\end{aligned}$$

because the JL-matrix $\mathbf{R}_f$ provides an 0.5-approximation. Since i is not checked, its scaling does not change, i.e. $\mathbf{V}_i^{(t_k)} = \mathbf{V}_i^{(t_{k-1})}$. We obtain

$$\left|\left\|e_j^\top\mathbf{V}^{(t_k)}\mathbf{A}\mathbf{B}^{(\mathrm{new})}\mathbf{S}_f\right\|_2 - \left\|e_j^\top\mathbf{V}^{(t_{k-1})}\mathbf{A}\mathbf{B}^{(\mathrm{old})}\mathbf{S}_f\right\|_2\right| \le \frac{0.05\epsilon\sqrt{z_i}}{\log T}.$$

Due to the inverse accuracy and (46), it holds that

$$\left\|e_j^\top\mathbf{V}^{(t_k)}\mathbf{A}\mathbf{B}^{(\mathrm{new})}\mathbf{S}_f\right\|_2^2 = \left\|e_j^\top\mathbf{V}^{(t_k)}\mathbf{A}\mathbf{M}^{(\mathrm{new})}\mathbf{A}^\top\mathbf{V}^{(t_k)}\mathbf{S}_f\right\|_2^2 \approx_{0.1\epsilon/\log T} \left\|e_j^\top\mathbf{P}(\mathbf{V}^{(t_k)}\mathbf{A})\right\|_2^2$$

Similarly, we have $\left\|e_j^\top\mathbf{V}^{(t_{k-1})}\mathbf{A}\mathbf{B}^{(\mathrm{old})}\mathbf{S}_f\right\|_2^2 \approx_{0.1\epsilon/\log T} \left\|e_j^\top\mathbf{P}(\mathbf{V}^{(t_{k-1})}\mathbf{A})\right\|_2^2$. We conclude that

$$\left|\left\|e_j^\top\mathbf{P}(\mathbf{V}^{(t_k)}\mathbf{A})\right\|_2^2 - \left\|e_j^\top\mathbf{P}(\mathbf{V}^{(t_{k-1})}\mathbf{A})\right\|_2^2\right| \le \frac{0.25\epsilon}{\log T}\left(\left\|e_j^\top\mathbf{P}(\mathbf{V}^{(t_k)}\mathbf{A})\right\|_2^2 + \overline{z}/n_i\right).$$

Summing over all $j \in \mathrm{G}_i$ and using $z_i = \overline{z}/n_i$ gives (49).

Stability argument. We first claim that Lemma C.15 in [VDBLL+21] also holds when the variables x_i, scalings $\mathbf{V}_i$, and other quantities are extended to multidimensional vectors and matrices. The argument is as follows.

First, Lemma C.15 in [VDBLL+21] relies on Lemma C.14. By examining the proof of Lemma C.14, we observe that the computations also extend to block-diagonal $\mathbf{V}$ and $\mathbf{V}'$ when their squares are spectrally close, i.e., $\mathbf{V}_i^2 \approx_{O(1)} (\mathbf{V}_i')^2$, without introducing any additional dimension or condition-number dependence. This condition is guaranteed by the assumptions of our lemma. We also replace the weighted norm of the change in $\mathbf{V}$ with the corresponding weighted block norm.

The remaining stabilizer argument follows the same structure as Lemma C.15, replacing coordinate leverage scores with block leverage scores and scalar relative changes with block relative operator norms.

The required conditions follow directly from the definitions of $\mathbf{V}$ and $\mathbf{F}$, together with the change condition in the lemma statement. This recovers the same asymptotic guarantee as Lemma C.15:

$$\left\|\mathbf{F}\mathbf{V}^{(t-2^l)}\mathbf{A}\left(\mathbf{M}^{(\text{new})}\mathbf{A}^\top\mathbf{V}^{(t)} - \mathbf{M}^{(\text{old})}\mathbf{A}^\top\mathbf{V}^{(t-2^l)}\right)\mathbf{S}_f\right\|_{\text{F}}^2 \lesssim 2^{2l} + \sum_{s=t-2^l}^{t}\sum_{\mathbf{V}_i^{(s)}\neq\mathbf{V}_i^{(s-1)}}\sigma(\mathbf{V}^{(s)}\mathbf{A})_i \quad (50)$$

Also, we replace the exact projection matrix at $\mathbf{V}^{(t)}$ and $\mathbf{V}^{(t-2^l)}$ with the approximations, and it only adds an additional $O(1)$ factor.

Runtime of VerifyBlocks. We analyze the running time spent by VerifyBlocks. For every single call to Query, it requires time to compute $\left\|e_j^\top\mathbf{V}\mathbf{A}\mathbf{H}_v\right\|_2^2$ for every $j \in \text{G}_i$ and every $i \in I$ needed to be checked. For any i, it takes $\widetilde{O}(\epsilon^{-2}n_i d) = \widetilde{O}(\epsilon^{-2}d)$ time when $n_i = O(1)$. Therefore we need to bound the cardinality of I by discussing three sources of block index $i \in I$. First, i is returned as a heavy block by heavy-hitters at the t-th call. Fix dyadic level l and interval $(t-2^l, t]$ for some $2^l|t$. If i is detected by D_l, there exists some $j \in \text{G}_i$ and $k \in [r]$ such that $0.02\epsilon/(r\log T) \le \left|e_j^\top\mathbf{F}\mathbf{Z}^{-1/2}\mathbf{V}^{(t-2^l)}\mathbf{A}\mathbf{H}_f e_k\right|$. Consequently, every returned block contributes at least $0.02^2\epsilon^2/(r_f^2\log^2 T)$ to the following total squared energy

$$\begin{aligned}\sum_{i\in[m]}\sum_{k\in[r]}\sum_{j\in\text{G}_i}\left|e_j^\top\mathbf{F}\mathbf{Z}^{-1/2}\mathbf{V}^{(t-2^l)}\mathbf{A}\mathbf{H}_f e_k\right|_2^2 &\le O(\left\|\mathbf{F}\mathbf{Z}^{-1/2}\mathbf{V}^{(t-2^l)}\mathbf{A}(\mathbf{B}^{(\text{new})} - \mathbf{B}^{(\text{old})})\mathbf{S}_f\right\|_{\text{F}}^2) \\ &\le \widetilde{O}\left(\overline{z}^{-1}\left(2^{2l} + \sum_{s=t-2^l}^{t}\sum_{\mathbf{V}_i^{(s)}\neq\mathbf{V}_i^{(s-1)}}\sigma(\mathbf{V}^{(s)}\mathbf{A})_i\right)\right).\end{aligned}$$

Summing over all $2^l|t$ and all $l = 0, \ldots, L$, the running time caused by first kind is bounded by

$$\widetilde{O}\left(Td + T^2 m\epsilon^{-4} + m\epsilon^{-4}\sum_{t=1}^{T}\sum_{\mathbf{V}_i^{(t)}\neq\mathbf{V}_i^{(t-1)}}\sigma(\mathbf{V}^{(t)}\mathbf{A})_i\right).$$

Second, i is added to I because its heavy-hitter scaling is changed after a call to Update. Any such call inserts i to every F_l and therefore creates L future block checks. This brings

$$\widetilde{O}\left(\sum_{t=1}^{T}\sum_{\mathbf{V}_i^{(t)}\neq\mathbf{V}_i^{(t-1)}} d\epsilon^{-2}\right)$$

total cost for T calls. For the third source, i is added to I because $\overline{\sigma}_i$ is reset at a previous call to QUERY. A block version of Lemma C.12 in [VDBLL+21] indicates the total weighted number of resets satisfies

$$\sum_{t=1}^{T}\sum_{i\in[m]}\left(\sigma(\mathbf{V}^{(t)}\mathbf{A})_i+\overline{z}\right)\mathbb{1}[\overline{\sigma}_i^{(t)}\neq\overline{\sigma}_i^{(t-1)}]\leq\widetilde{O}\left(\epsilon^{-1}\sum_{t=1}^{T}\sum_{\mathbf{V}_i^{(t)}\neq\mathbf{V}_i^{(t-1)}}\sigma(\mathbf{V}^{(t)}\mathbf{A})_i\right).$$

Lemma C.12 can be extended to the block setting because the change in the Gram matrix is supported on the updated blocks, and the spectral comparability of consecutive block squares bounds the total variation by their block leverage scores. The remaining argument can be unchanged, without additional dimension or condition number dependence. Then the running cost due to the third source is bounded by

$$\widetilde{O}\left(m\epsilon^{-3}\sum_{t=1}^{T}\sum_{\mathbf{V}_i^{(t)}\neq\mathbf{V}_i^{(t-1)}}\sigma(\mathbf{V}^{(t)}\mathbf{A})_i\right).$$

In addition, it requires $\widetilde{O}(\epsilon^{-4}Td^2+\epsilon^{-2}TT_{\text{inverse}})$ time to compute $\mathbf{H}_v$ for T calls.

Runtime of FindBlocks. We first analyze the cost of formalizing $\mathbf{H}_f$. Fix l and consider the t-th call with $2^l|t$. Let $|S_f|$ denote the number of blocks preserved by $\mathbf{S}_f$. We only materialize $\mathbf{A}_i^\top\mathbf{V}_i^{(t)}$ for the blocks preserved by $\mathbf{S}_f$, and only apply the inverse $\mathbf{M}^{(\text{new})}$ on $\mathbf{A}^\top\mathbf{V}^{(t)}\mathbf{S}_f\mathbf{R}_f\in\mathbb{R}^{d\times r}$, the cost is $\widetilde{O}(T_{\text{inverse}}+|S_f|d)$ when $n_i=O(1)$ and $r=\widetilde{O}(1)$. We will discuss inverse cost in the end. It remains to bound $|S_f|$. While all the blocks in F_l are deterministically preserved by $\mathbf{S}_f$, w.h.p. $\widetilde{O}(d\epsilon^{-2})$ other blocks are preserved by sampling. Recall that the i-th block is inserted into some F_l only when D_l.UPDATE is called or $\overline{\sigma}_i$ is recently reset. We again charge the cost to UPDATE when it is called. As a consequence, the remaining running time of computing $\mathbf{H}_f$ over T calls and all dyadic level is bounded by

$$\widetilde{O}\left(Td^2\epsilon^{-2}+m\epsilon^{-2}\sum_{t=1}^{T}\sum_{\mathbf{V}_i^{(t)}\neq\mathbf{V}_i^{(t-1)}}\sigma(\mathbf{V}^{(t)}\mathbf{A})_i\right).$$

We next analyze the cost of heavy-hitters. For dyadic level l and interval $(t-2^l,t]$ with $2^l|t$, it takes time

$$\widetilde{O}\left(d\left\|\mathbf{F}\mathbf{Z}^{-1/2}\mathbf{V}^{(t-2^l)}\mathbf{A}\mathbf{H}_f\right\|_{\mathrm{F}}^2\epsilon^{-2}+d\right)\leq\ \widetilde{O}\left(d+m\epsilon^{-2}2^{2l}+m\epsilon^{-2}\sum_{s=t-2^l+1}^{t}\sum_{\mathbf{V}_i^{(s)}\neq\mathbf{V}_i^{(s-1)}}\sigma(\mathbf{V}^{(s)}\mathbf{A})_i\right).$$

Summing over all T calls and all levels l, we get

$$\widetilde{O}\left(Td+m\epsilon^{-2}T^2+m\epsilon^{-2}\sum_{t=1}^{T}\sum_{\mathbf{V}_i^{(t)}\neq\mathbf{V}_i^{(t-1)}}\sigma(\mathbf{V}^{(t)}\mathbf{A})_i\right).$$

In addition, it requires $\widetilde{O}(TT_{\text{inverse}})$ time to compute the inverse.

Runtime analysis. Combining the bounds for VerifyBlocks and FindBlocks, the total running time over T calls to Query and all Scale is

$$\widetilde{O}\left(Td^2\epsilon^{-4} + T^2m\epsilon^{-4} + T\epsilon^{-2}T_{\text{inverse}} + \sum_{t=1}^{T}\sum_{\mathbf{V}_i^{(t)}\neq\mathbf{V}_i^{(t-1)}}\left(d\epsilon^{-2} + m\epsilon^{-4}\sigma(\mathbf{V}^{(t)}\mathbf{A})_i\right)\right).$$

Here we charge all costs caused by updating the scaling of block i to the corresponding call to Scale. The costs associated with resets are charged to the same scaling updates above. All remaining costs are charged to Query. Consequently, each call to Scale takes amortized time

$$\widetilde{O}\left(d\epsilon^{-2} + m\epsilon^{-4}\sigma(\mathbf{V}^{(t)}\mathbf{A})_i\right).$$

Each call to Query has

$$\widetilde{O}\left(d^2\epsilon^{-4} + Tm\epsilon^{-4} + \epsilon^{-2}T_{\text{inverse}}\right).$$

□

C.4 Samping and Inverse Maintenance

Proof of Lemma 47. We follow the block sampling data structure of [LPT+25, Lemma 6.2], which supports sampling proportional to $\|\mathbf{W}_i\mathbf{A}_ih\|_2^2$ under dynamic block scaling updates. We additionally maintain the weights τ_i and their sum in a weighted sampling tree, supporting individual weight updates and sampling proportional to τ_i in $\widetilde{O}(1)$ time.

To implement Sample, we combine three distributions: norm-proportional sampling using [LPT+25, Lemma 6.2], uniform block sampling, and sampling proportional to τ_i, with respective mixture probabilities $m/(K\sqrt{d})$, $m/(K\sqrt{d})$, and $\sum_i \tau_i/K$. This gives exactly the distribution required by the Sampling Scheme lemma. The additional weighted sampling tree does not change the stated running times. Moreover, the exact rejection correction in the original sampler preserves the target distribution under adaptive queries, while the additional sampling operations use fresh randomness. □

Proof of Lemma 48. We slightly modify Algorithms 8 and 9 in [vdBLSS20] to work with block-diagonal positive definite matrices. A call to Update only changes the stored block weight and leverage hint, which takes $\widetilde{O}(1)$ time since $n_i = O(1)$. At each call to Solve, we first synchronize the maintained inverse using the block version of Algorithm 8. Hence, the cost of inverse maintenance is charged to Solve.

We first discuss the modifications to Algorithm 8. We use $\mathbf{W}^{(\text{alg})}$, $\tau^{(\text{alg})}$, $\mathbf{V}$, and $\Psi = (\mathbf{A}^\top\mathbf{V}\mathbf{A})^{-1}$ for the cached weights, cached leverage hints, sampled weights, and maintained inverse, respectively. In Lines 9 and 19, we sample each entire block independently with

$$p_i = \min\{1, \gamma\epsilon^{-2}\tau_i^{(\text{alg})}\}, \qquad \mathbf{V}_i = \frac{\chi_i}{p_i}\mathbf{W}_i^{(\text{alg})}, \quad \chi_i \sim \text{Bernoulli}(p_i).$$

In Line 12, we replace the scalar relative-weight error by

$$y_i = \frac{8}{\epsilon}\left\|\mathbf{W}_i - \mathbf{W}_i^{(\text{alg})}\right\|_{(\mathbf{W}_i^{(\text{alg})})^{-1}},$$

while keeping the leverage hint error unchanged. We retain the sorting and dyadic threshold rules in Lines 13–19, using $\max\{\tau_i, \tau_i^{(\mathrm{alg})}\}$ as the mass of both records of block i to account for the old and new samples. The hints may be capped at n_i since $l_i(\mathbf{W}) \le n_i$. Selecting either record refreshes both cached quantities and resamples the corresponding block. In Line 20, each refreshed block contributes rank $O(n_i) = O(1)$, so the same batched Woodbury procedure applies.

For correctness, the update rule guarantees $\mathbf{W}_i^{(\mathrm{alg})} \approx_{O(\epsilon)} \mathbf{W}_i$ and $\tau_i^{(\mathrm{alg})} \approx_{O(1)} \tau_i$. Writing $\mathbf{K}^{(\mathrm{alg})} = \mathbf{A}^\top \mathbf{W}^{(\mathrm{alg})} \mathbf{A}$, we have

$$\mathbf{Q}_i = (\mathbf{K}^{(\mathrm{alg})})^{-1/2} \mathbf{A}_i^\top \mathbf{W}_i^{(\mathrm{alg})} \mathbf{A}_i (\mathbf{K}^{(\mathrm{alg})})^{-1/2},$$
$$\sum_i \mathbf{Q}_i = \mathbf{I}, \qquad \|\mathbf{Q}_i\|_{\mathrm{op}} \le \mathrm{tr}(\mathbf{Q}_i) = l_i(\mathbf{W}^{(\mathrm{alg})}) \lesssim \tau_i^{(\mathrm{alg})}.$$

Thus the matrix Chernoff bound gives

$$\mathbf{A}^\top \mathbf{V} \mathbf{A} \approx_{O(\epsilon)} \mathbf{A}^\top \mathbf{W}^{(\mathrm{alg})} \mathbf{A} \approx_{O(\epsilon)} \mathbf{K}(\mathbf{W}),$$

with $\widetilde{O}(\epsilon^{-2} d)$ sampled blocks.

We now discuss the running time. The proof follows [vdBLSS20, Theorem 10], where the scalar relative-weight increment bound in (8.5) is replaced by

$$\left| \|\widehat{\mathbf{W}}_i^{(t+1)} - \widehat{\mathbf{W}}_i^{(s)}\|_{(\widehat{\mathbf{W}}_i^{(s)})^{-1}} - \|\widehat{\mathbf{W}}_i^{(t)} - \widehat{\mathbf{W}}_i^{(s)}\|_{(\widehat{\mathbf{W}}_i^{(s)})^{-1}} \right|$$
$$\le \|\widehat{\mathbf{W}}_i^{(t+1)} - \widehat{\mathbf{W}}_i^{(t)}\|_{(\widehat{\mathbf{W}}_i^{(s)})^{-1}} \lesssim \|\widehat{\mathbf{W}}_i^{(t+1)} - \widehat{\mathbf{W}}_i^{(t)}\|_{(\widehat{\mathbf{W}}_i^{(t)})^{-1}},$$

whenever block i has not been refreshed since iteration s. The $\widehat{\tau}$-weighted stability condition gives the same charging argument as (8.4)–(8.5), while $\max\{\tau_i, \tau_i^{(\mathrm{alg})}\}$ accounts for both the old and new samples. Since $n_i = O(1)$, the remaining Woodbury amortization is unchanged. Thus synchronizing the inverse takes $\widetilde{O}(n + \epsilon^{-2}(d^2 + d^{\omega - 1/2}))$ amortized time per call to Solve.

Finally, we modify Algorithm 9 by replacing fresh row sampling with independent block sampling. Since $\mathbf{W}_i' \approx_1 \mathbf{W}_i$, the current leverage hints remain constant-factor upper bounds for $l_i(\mathbf{W}')$. For a fresh sampled block-diagonal matrix μ of $\mathbf{W}'$, we have

$$\mathbb{E}[\mathbf{A}^\top \mu \mathbf{A}] = \mathbf{K}(\mathbf{W}'), \qquad \mathrm{Cov}(\mathbf{A}^\top \mu^{1/2} g \mid \mu) = \mathbf{A}^\top \mu \mathbf{A}, \quad g \sim N(0, \mathbf{I}).$$

Hence, the same spectral concentration, iterative refinement, and protected-solve arguments apply to the block samples. They give the required ϵ-accurate solution with high probability in $\widetilde{O}(n + \epsilon^{-2} d^2)$ time, which is absorbed by the stated running time. □

D Remaining Proofs for Section 6

D.1 Initial and Final Processing

Proof of Lemma 49. We verify the first point. It is easy to verify $\begin{bmatrix} x^{(\mathrm{init})} \\ z^{(\mathrm{init})} \end{bmatrix}$ and $\begin{bmatrix} s^{(\mathrm{init})} \\ r^{(\mathrm{init})} \end{bmatrix}$ are primal and dual feasible. Now fix $i \in [m]$. For any vector $v \in \mathbb{R}^{n_i}$ with $\|\nabla^2 \phi(x_i^{(\mathrm{init})})^{\frac{1}{2}} v\|_2 \le 1$, it holds

that $\left\|x_i^{(\text{init})} \pm v\right\|_2 \leq R$ by an application of Theorem 4.1.7 in [Nes13]. Consequently, it holds that $\nabla^2\phi(x_i^{(\text{init})})^{-1} \preceq R^2\mathbf{I}$. We compute

$$\left\|\frac{s_i^{(\text{init})}}{\rho\tau\left(x^{(\text{init})}\right)_i} + \nabla\phi_i(x_i^{(\text{init})})\right\|_{\nabla^2\phi(x_i^{(\text{init})})^{-1}} \leq \frac{1}{\rho\tau\left(x^{(\text{init})}\right)_i}\left\|\nabla^2\phi_i\left(x_i^{(\text{init})}\right)^{-\frac{1}{2}}c_i\right\|_2 \leq \frac{mR}{\rho d}\left\|c_i\right\|_2 \leq \delta/2$$

due to $\|c\|_{2,\infty} \leq L$, $\tau_i \geq d/m$, $\rho = \frac{5RmL}{d\delta^2}$, and also $\nabla\phi_i(x_i^{(\text{init})}) = 0$. For θ, it holds that $f'(1) = 0$ and also

$$\left|h''(z^{(\text{init})})^{-\frac{1}{2}}\left(\frac{r^{(\text{init})}}{\rho\tau([x^{(\text{init})}, z^{(\text{init})}])_{m+1}} + h'(z^{(\text{init})})\right)\right| = \left|\frac{|h''(1)|^{-\frac{1}{2}}LR}{\rho\tau([x^{(\text{init})}, z^{(\text{init})}])_{m+1}}\right| \leq \frac{2mLR}{\rho\delta d} \leq \delta/2.$$

As a consequence of Lemma 8, the triplet is δ-centered.

Now we prove the second point. It is obvious that $x \in K$, therefore feasible for the original program. From the optimality condition it holds that

$$\min_{\widetilde{x}\in K\times[0,2],\ \widetilde{\mathbf{A}}^\top\widetilde{x}=\widetilde{b}} \widetilde{c}^\top\widetilde{x} \leq \min_{x\in K,\mathbf{A}^\top x=b} c^\top x + LRz/\delta.$$

We combine the above inequality with the condition $c^\top x + LRz/\delta \leq \min_{\widetilde{x}\in K\times(0,2),\ \widetilde{\mathbf{A}}^\top\widetilde{x}=\widetilde{b}} \widetilde{c}^\top\widetilde{x}$ and obtain immediately $c^\top x \leq \min_{x\in K,\mathbf{A}^\top x=b} c^\top x \leq c^\top x + \delta$.

Now we analyze approximate feasibility. It holds that $\left|c^\top x\right| \leq mLR$ and $\min_{x\in K,\mathbf{A}^\top x=b} c^\top x \leq mLR$, we estimate

$$LR/\delta \cdot z \leq -c^\top x + \min_{x\in K,\mathbf{A}^\top x=b} c^\top x + \delta \leq 2mLR + \delta.$$

This implies $z \leq 2m\delta + \delta^2/(LR)$. Since $\widetilde{x}$ is strictly feasible for the extended program, we have $\mathbf{A}^\top x - b = z(\mathbf{A}^\top x^{(\text{init})} - b)$. As a result

$$\left\|\mathbf{A}^\top x - b\right\|_\infty \leq \left\|\mathbf{A}^\top x - b\right\|_1 \leq z\left\|\mathbf{A}^\top x^{(\text{init})} - b\right\|_1 \leq 2mRM\left(2m\delta + \frac{\delta^2}{LR}\right).$$

by inserting the bound for $\mathbf{A}$ and x again. □

Proof of 50. We define the output pair as

$$x^{(\text{fi})} = x - \overline{\mathbf{T}}^{-1}\nabla^2\Phi(x)^{-1}\mathbf{A}(\mathbf{A}^\top\overline{\mathbf{T}}^{-1}\nabla^2\Phi(x)^{-1}\mathbf{A})^{-1}(\mathbf{A}^\top x - b), \qquad s^{(\text{fi})} = s.$$

It is obvious that $x^{(\text{fi})}$ and $s^{(\text{fi})}$ are primal and dual feasible. Define $\mu_i^{(\text{fi})} = \frac{s_i^{(\text{fi})}}{\rho\tau_i(x)} + \nabla\phi_i(x_i^{(\text{fi})})$ for any $i \in [m]$. We now show $\left\|\mu_i^{(\text{fi})}\right\|_{\nabla^2\phi_i(x_i^{(\text{fi})})^{-1}}$ is bounded by $O(\epsilon)$. First, we apply the same argument to $x^{(\text{fi})} - x$ and obtain

$$\left\|\nabla^2\Phi(x)^{\frac{1}{2}}(x^{(\text{fi})} - x)\right\|_{2,\infty} \leq \left\|\mathbf{A}^\top x - b\right\|_{(\mathbf{A}^\top\mathbf{T}^{-1}\nabla^2\Phi(x)^{-1}\mathbf{A})^{-1}} \leq \frac{\alpha\beta}{2C_{\text{norm}}}.$$

As a consequence, it holds that

$$\begin{aligned}\left\|\mu_i^{(\mathrm{fi})}\right\|_{\nabla^2\phi_i(x_i^{(\mathrm{fi})})^{-1}} \leq & 2\left\|\frac{s_i^{(\mathrm{fi})}}{\rho\tau_i(x)}+\nabla\phi_i(x_i^{(\mathrm{fi})})\right\|_{\nabla^2\phi_i(x_i)^{-1}} \\ \leq & 2\left\|\frac{s_i}{\rho\tau_i(x)}+\nabla\phi_i(x_i)\right\|_{\nabla^2\phi_i(x_i)^{-1}}+2\left\|\nabla\phi_i(x_i^{(\mathrm{fi})})-\nabla\phi_i(x_i)\right\|_{\nabla^2\phi_i(x_i)^{-1}} \\ \leq & 2\left\|\frac{s_i}{\rho\tau_i(x)}+\nabla\phi_i(x_i)\right\|_{\nabla^2\phi_i(x_i)^{-1}}+4\left\|\nabla^2\phi_i(x_i)^{\frac{1}{2}}(x_i^{(\mathrm{fi})}-x_i)\right\|_2 \leq 6\epsilon\end{aligned}$$

where we use self-concordance, Lemma 3, Lemma 8, and the fact that (x, s, ρ) is ϵ-centered.

It remains to show the bound the suboptimality. Let $x^* = \operatorname{argmin}_{x\in K, \mathbf{A}^\top x=b} c^\top x$ and $\overline{x} = \frac{x^{(\mathrm{fi})}+x^*}{2}$. Theorem 4.1.7 in [Nes13] indicates $\langle\nabla\phi_i(\overline{x}_i), x_i^* - \overline{x}_i\rangle \leq 1$. Since $\overline{x} - x^* = x^{(\mathrm{fi})} - \overline{x}$, we also have $\langle\nabla\phi_i(\overline{x}_i), \overline{x}_i - x_i^{(\mathrm{fi})}\rangle \leq 1$. It holds that

$$\begin{aligned}&\left\langle\nabla\phi_i(\overline{x}_i), \overline{x}_i - x_i^{(\mathrm{fi})}\right\rangle \\ =&\left\langle\nabla\phi_i(\overline{x}_i)-\nabla\phi_i(x_i^{(\mathrm{fi})}), \overline{x}_i - x_i^{(\mathrm{fi})}\right\rangle+\left\langle\mu_i^{(\mathrm{fi})}-\frac{s_i^{(\mathrm{fi})}}{\rho\tau_i(x)}, \overline{x}_i - x_i^{(\mathrm{fi})}\right\rangle \\ \geq&\frac{1}{2\rho\tau_i(x)}\left\langle s_i, x_i^{(\mathrm{fi})}-x_i^*\right\rangle+\frac{\left\|\nabla^2\phi_i(x_i^{(\mathrm{fi})})^{\frac{1}{2}}(x_i^*-x_i^{(\mathrm{fi})})\right\|_2^2}{4+2\left\|\nabla^2\phi_i(x_i^{(\mathrm{fi})})^{\frac{1}{2}}(x_i^*-x_i^{(\mathrm{fi})})\right\|_2}-\frac{1}{2}\left\|\mu_i^{(\mathrm{fi})}\right\|_{\nabla^2\phi_i(x_i^{(\mathrm{fi})})^{-1}}\left\|\nabla^2\phi_i(x_i^{(\mathrm{fi})})^{\frac{1}{2}}(x_i^*-x_i^{(\mathrm{fi})})\right\|_2 \\ \geq&\frac{1}{2\rho\tau_i(x)}\left\langle s_i, x_i^{(\mathrm{fi})}-x_i^*\right\rangle-\frac{\left\|\nabla^2\phi_i(x_i^{(\mathrm{fi})})^{\frac{1}{2}}(x_i^*-x_i^{(\mathrm{fi})})\right\|_2}{2+\left\|\nabla^2\phi_i(x_i^{(\mathrm{fi})})^{\frac{1}{2}}(x_i^*-x_i^{(\mathrm{fi})})\right\|_2}\end{aligned}$$

due to Lemma 3 and $\left\|\mu_i^{(\mathrm{fi})}\right\|_{\nabla^2\phi_i(x_i^{(\mathrm{fi})})^{-1}} \leq O(\epsilon) \leq 1$. As a consequence, we compute

$$\begin{aligned}c^\top\left(x^{(\mathrm{fi})}-x^*\right)=\left\langle\mathbf{A}y+s, x^{(\mathrm{fi})}-x^*\right\rangle=&\sum_{i=1}^m\left\langle s_i, x_i^{(\mathrm{fi})}-x_i^*\right\rangle \\ \leq& 2\rho\sum_{i=1}^m\frac{\tau_i(x)\left\|\nabla^2\phi_i(x_i^{(\mathrm{fi})})^{\frac{1}{2}}(x_i^*-x_i^{(\mathrm{fi})})\right\|_2}{2+\left\|\nabla^2\phi_i(x_i^{(\mathrm{fi})})^{\frac{1}{2}}(x_i^*-x_i^{(\mathrm{fi})})\right\|_2}+2\rho\sum_{i=1}^m\tau_i(x) \\ \leq& 4\rho\sum_{i=1}^m\tau_i(x)=8d\rho\end{aligned}$$

by combining the above inequality and $\langle\nabla\phi_i(\overline{x}_i), \overline{x}_i - x_i^{(\mathrm{fi})}\rangle \leq 1$. □

D.2 Proof of Nearby Sequence

Proof of Lemma 51. Let $\widehat{x}^{(1)} = x^{(1)}$. For the k-th iteration, define the stability potential Ψ_{nb} to measure the discrepancy between $\widehat{x}^{(k)}$ and $x^{(k)}$

$$\Psi_{\mathrm{nb}}(\widehat{x}^{(k)}, x^{(k)}) := \sum_{i=1} e^{\lambda_{\mathrm{nb}}\gamma_{\mathrm{nb},i}^{(k)}}, \qquad \gamma_{\mathrm{nb},i}^{(k)} := \left\|\widehat{x}_i^{(k)} - x_i^{(k)}\right\|_{\nabla^2\phi_i(x_i^{(k)})}^2,$$

where $\lambda_{\rm nb} = 20\log(mT)/\alpha_{\rm nb}^2$. We use abbreviation $\Psi_{\rm nb}^{(k)} = \Psi_{\rm nb}(\widehat{x}^{(k)}, x^{(k)})$ when there is no ambiguity. We claim the following fact is true for any $k > 1$ and postpone its proof for clarity:

$$\mathbb{E}\left[\Psi_{\rm nb}(\widehat{x}^{(k+1)}, x^{(k+1)})\right] \le \left(1 - \frac{\theta\beta\log(mT)}{\sqrt{d}}\right)\Psi_{\rm nb}(\widehat{x}^{(k)}, x^{(k)}) + m^2. \tag{51}$$

Since $\Psi_{\rm nb}(\widehat{x}^{(1)}, x^{(1)}) = m$, it immediately follows that $\mathbb{E}\left[\Psi_{\rm nb}(\widehat{x}^{(k)}, x^{(k)})\right] \le m^3$ is true for any k. With Markov's inequality, we conclude with probability $1 - 1/\mathrm{poly}(m)$, $\Psi_{\rm nb}(\widehat{x}^{(k)}, x^{(k)}) \le m^{15}$. By the choice of $\lambda_{\rm nb}$ and $\gamma_{\rm nb}$, we establish the first point

$$\left\|\nabla^2\Phi(x^{(k)})^{\frac{1}{2}}(\widehat{x}^{(k)} - x^{(k)})\right\|_{2,\infty} \le \alpha_{\rm nb}.$$

where $\beta_{\rm nb} = C\alpha_{\rm nb}\beta$. This implies $\nabla^2\Phi(\widehat{x}^{(k)}) \approx_{O(\alpha_{\rm nb})} \nabla^2\Phi(x^{(k)})$ due to Point 2 in Lemma 3. Consequently,

$$\left\|\nabla^2\Phi(\widehat{x}^{(k)})^{\frac{1}{2}}(\widehat{x}^{(k)} - x^{(k)})\right\|_{2,\infty} \le 2\alpha_{\rm nb}$$

also holds.

Now suppose all the three conditions in the lemma statement are true at iteration k. An application of Corollary 25 implies $\tau(x^{(k)}) \approx_{O(\alpha_{\rm nb})} \tau(\widehat{x}^{(k)})$. The main job is therefore to establish the one-step convergence of $\Psi_{\rm nb}$ in (51). We will choose $\widehat{x}^{(k+1)}$ using the gradient step of $\Psi_{\rm nb}$. Let

$$g_{\rm nb}^{(k)} = -\beta_{\rm nb}\left(\nabla^2\Phi(x^{(k)})^{-\frac{1}{2}}\nabla_{\widehat{x}}\Psi(x^k, \widehat{x}^k)\right)^{\flat(\tau(\widehat{x}^k))}.$$

The update in $\widehat{x}$ is defined by

$$\delta_{\widehat{x}} = \widehat{x}^{(k+1)} - \widehat{x}^{(k)} := \mathbb{E}[x^{(k+1)} - x^{(k)}] + \nabla^2\Phi(x^{(k)})^{-\frac{1}{2}}g_{\rm nb}^{(k)} = \mathbb{E}[\delta_x] + \nabla^2\Phi(x^{(k)})^{-\frac{1}{2}}g_{\rm nb}^{(k)}. \tag{52}$$

As a direct consequence of the above expansion, we obtain the third point

$$\left\|\nabla^2\Phi(x^{(k)})^{\frac{1}{2}}(\widehat{x}^{(k+1)} - \widehat{x}^{(k)})\right\|_{2,\tau(x^{(k)})+\infty} \le \left\|\nabla^2\Phi(x^{(k)})^{\frac{1}{2}}\mathbb{E}[\delta_x]\right\|_{2,\tau(x^{(k)})+\infty} + \left\|g_{\rm nb}^{(k)}\right\|_{2,\tau(x^{(k)})+\infty} \le 2\beta$$

due to $\tau(x^{(k)}) \approx_{O(\alpha_{\rm nb})} \tau(\widehat{x}^{(k)})$ and norm bound $\left\|g_{\rm nb}^k\right\|_{\tau(\widehat{x}^{(k)})+\infty} = \beta_{\rm nb} \lesssim \beta$.

A preliminary step before we establish (51) is to control the maximal change in $\gamma_{\rm nb}$. The following expansion holds for $\delta_{\gamma_{{\rm nb},i}^{(k)}}$:

$$\begin{aligned}\delta_{\gamma_{{\rm nb},i}^{(k)}} = \gamma_{{\rm nb},i}^{(k+1)} - \gamma_{{\rm nb},i}^{(k)} = &\left\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\delta_{\widehat{x},i} - \delta_{x,i})\right\|_2^2 + 2\left\langle\delta_{\widehat{x},i} - \delta_{x,i}, \nabla^2\phi_i(x_i^{(k)})(\widehat{x}_i^{(k)} - x_i^{(k)})\right\rangle \\ &+ \left\langle\widehat{x}_i^{(k+1)} - x_i^{(k+1)}, (\nabla^2\phi_i(x_i^{(k+1)}) - \nabla^2\phi_i(x_i^{(k)}))(\widehat{x}_i^{(k+1)} - x_i^{(k+1)})\right\rangle.\end{aligned} \tag{53}$$

It is necessary to bound each term in the RHS of (53). For the first term, we consider a stochastic-deterministic splitting using the identity in (52)

$$\left\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\delta_{\widehat{x},i} - \delta_{x,i})\right\|_2^2 \le 2\left\|g_{{\rm nb},i}^{(k)}\right\|_2^2 + 2\left\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\delta_{x,i} - \mathbb{E}[\delta_{x,i}])\right\|_2^2$$

which in addition relies on the fact $(a+b)^2 \le 2a^2 + 2b^2$. In particular, the stochastic part admits the following bound

$$\left\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\delta_{x,i} - \mathbb{E}[\delta_{x,i}])\right\|_2 \le \left\|\nabla^2\Phi(x^{(k)})^{\frac{1}{2}}(\delta_x - \mathbb{E}[\delta_x])\right\|_{2,\infty}$$

$$\leq \left\|\delta_r - \mathbf{R}\delta_r\right\|_{2,\infty} \leq \beta\sqrt{\frac{d}{m}}\frac{\|\delta_r\|_2}{C_{\text{var}}^2} \leq \frac{\beta\|\delta_r\|_{2,\tau}}{C_{\text{var}}^2} \lesssim \frac{\beta^2}{C_{\text{var}}^2} \tag{54}$$

where the second step is from the definition of δ_x in Algorithm 1, the third uses the maximal property of $\mathbf{R}$ in Definition 19. Also, we use

$$\|\delta_r\|_{\tau+\infty} \leq \|\delta_1\|_{\tau+\infty} + \|\delta_2\|_{\tau+\infty} \lesssim \beta,$$

as a consequence of Lemma 26. The deterministic part is simply controlled through from the norm bound $\left\|g_{\text{nb}}^{(k)}\right\|_{2,\tau(\widehat{x}^{(k)})+\infty} = \beta_{\text{nb}}$. Therefore we obtain

$$\left\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\delta_{\widehat{x},i} - \delta_{x,i})\right\|_2^2 \lesssim \beta_{\text{nb}}^2 \tag{55}$$

using $C_{\text{Var}}^2 \geq \beta/\alpha_{\text{nb}}$. It requires an additional mixed norm bound for future use. From the Variance property of the block-diagonal sampling, it holds that

$$\mathbb{E}\left[\|\nabla^2\phi_i(x_i^{(k)})^{1/2}(\delta_{x,i} - \mathbb{E}[\delta_{x,i}])\|_2^2\right] \lesssim \frac{\beta^2}{C_{\text{var}}^4}\sqrt{\frac{d}{m}}\,\|\delta_r\|_{2,\tau}\,\|\delta_{r,i}\|_2 \lesssim \frac{\beta^3}{C_{\text{norm}}C_{\text{var}}^4}\,\|\delta_{r,i}\|_2\,.$$

As a consequence,

$$\left\|\left(\mathbb{E}[\|\nabla^2\phi_i(x_i^{(k)})^{1/2}(\delta_{x,i} - \mathbb{E}[\delta_{x,i}])\|_2^2]\right)_{i\in[m]}\right\|_{\tau(\widehat{x}^{(k)})+\infty} \lesssim \frac{\beta_{\text{nb}}^2}{C^2} + \frac{\beta^3}{C_{\text{var}}^4}\,\|\delta_r\|_{2,\tau} \lesssim \beta_{\text{nb}}^2/C^2$$

due to our assumption $C_{\text{var}}^2 \geq \beta/\alpha_{\text{nb}}$.

We compute the upper bound for the second term in (53)

$$\begin{aligned}&\left|\langle\delta_{\widehat{x},i} - \delta_{x,i}, \nabla^2\phi_i(x_i^{(k)})(\widehat{x}_i^{(k)} - x_i^{(k)})\rangle\right| \\ &\quad\leq \left\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\widehat{x}_i^{(k)} - x_i^{(k)})\right\|_2\left(\left\|g_{\text{nb},i}^{(k)}\right\|_2 + \left\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\delta_{x,i} - \mathbb{E}[\delta_{x,i}])\right\|_2\right) \lesssim \alpha_{\text{nb}}\beta_{\text{nb}},\end{aligned} \tag{56}$$

where the final line comes from the induction $\left\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\widehat{x}_i^{(k)} - x_i^{(k)})\right\|_2 \leq \alpha_{\text{nb}}$ and (54). For the last term in (53), it holds that

$$\begin{aligned}&\left|\langle\widehat{x}_i^{(k+1)} - x_i^{(k+1)}, (\nabla^2\phi_i(x_i^{(k+1)}) - \nabla^2\phi_i(x_i^{(k)}))(\widehat{x}_i^{(k+1)} - x_i^{(k+1)})\rangle\right| \\ \leq&\left\|\nabla^2\phi_i(x_i^{(k)})^{-\frac{1}{2}}\nabla^2\phi_i(x_i^{(k+1)})\nabla^2\phi_i(x_i^{(k)})^{-\frac{1}{2}} - \mathbf{I})\right\|_{\text{op}}\|\widehat{x}_i^{(k+1)} - x_i^{(k+1)}\|_{\nabla^2\phi_i(x_i^{(k)})}^2 \\ \leq&2\left\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}\delta_{x,i}\right\|_2\left(\left\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\widehat{x}_i^{(k)} - x_i^{(k)})\right\|_2^2 + \left\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\delta_{\widehat{x},i} - \delta_{x,i})\right\|_2^2\right) \\ \lesssim&\alpha_{\text{nb}}^2\beta + \beta\beta_{\text{nb}}^2\end{aligned} \tag{57}$$

where the second inequality comes from self-concordance, Lemma 3, $(a+b)^2 \leq 2a^2 + 2b^2$, and the final inequality is due to Lemma 26, the induction, and the above bound in (55). As a direct consequence, we obtain

$$\|\delta_{\gamma_{\text{nb}}}\|_\infty \lesssim \alpha_{\text{nb}}\beta_{\text{nb}}/C.$$

using coefficient relationship $\beta_{\text{nb}} = C\alpha_{\text{nb}}\beta$.

We proceed to establish the one-step descent of the stability potential in (51). Similar to Lemma 34, we apply Lemma 4.10 in [LPT+25] to the update of each $\gamma_{\mathrm{nb},i}^{(k)}$ and obtain

$$\mathbb{E}\big[\Psi_{\mathrm{nb}}^{(k+1)}\big] - \Psi_{\mathrm{nb}}^{(k)} \le \sum_{i=1}^{m} \psi_{\mathrm{nb}}'(\gamma_{\mathrm{nb},i}^{(k)})\mathbb{E}\left[\delta_{\gamma_{\mathrm{nb},i}}\right] + \lambda_{\mathrm{nb}} \exp\left(\lambda_{\mathrm{nb}}\big\|\delta_{\gamma_{\mathrm{nb}}}\big\|_\infty\right) \sum_{i=1}^{m} \psi_{\mathrm{nb}}'(\gamma_{\mathrm{nb},i}^{(k)})\mathbb{E}\left[\delta_{\gamma_{\mathrm{nb},i}}^2\right] \tag{58}$$

where we use conditional expectation over past iterates and the fact $\psi_{\mathrm{nb}}''(\cdot) = \lambda_{\mathrm{nb}}\psi_{\mathrm{nb}}'(\cdot)$. We expand the linear parts in the above RHS using the $\delta_{\gamma_{\mathrm{nb}}}$ expansion in (53) and bound each term in a separate style. The first term in (53) leads to the following computation

$$\begin{aligned}
&\sum_{i=1} \psi_{\mathrm{nb}}'(\gamma_{\mathrm{nb},i}^{(k)})\mathbb{E}\big[\big\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\delta_{\widehat{x},i} - \delta_{x,i})\big\|_2^2\big] \\
=&\sum_{i=1}^{m} \psi_{\mathrm{nb}}'(\gamma_{\mathrm{nb},i}^{(k)})\mathbb{E}\left[\big\|g_{\mathrm{nb},i}^{(k)}\big\|_2^2 + \big\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\delta_{x,i} - \mathbb{E}[\delta_{x,i}])\big\|_2^2 + 2\big\langle g_{\mathrm{nb},i}^{(k)}, \nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\delta_{x,i} - \mathbb{E}[\delta_{x,i}]\big\rangle\right] \\
=&\sum_{i=1}^{m} \psi_{\mathrm{nb}}'(\gamma_{\mathrm{nb},i}^{(k)})\mathbb{E}\left[\big\|g_{\mathrm{nb},i}^{(k)}\big\|_2^2 + \big\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\delta_{x,i} - \mathbb{E}[\delta_{x,i}])\big\|_2^2\right] \\
\le&\big\|\nabla_\gamma \Psi_{\mathrm{nb}}^{(k)}\big\|^*_{\tau(\widehat{x}^{(k)})+\infty}\left(2\beta_{\mathrm{nb}}\big\|g_{\mathrm{nb}}^{(k)}\big\|_{2,\tau(\widehat{x}^{(k)})+\infty} + \big\|\big(\mathbb{E}[\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\delta_{x,i} - \mathbb{E}[\delta_{x,i}])\|_2^2]\big)_{i\in[m]}\big\|_{\tau(\widehat{x}^{(k)})+\infty}\right) \\
\lesssim&\beta_{\mathrm{nb}}^2\big\|\nabla_\gamma \Psi_{\mathrm{nb}}^{(k)}\big\|^*_{\tau(\widehat{x}^{(k)})+\infty}
\end{aligned}$$

where we use the norm bound of $g_{\mathrm{nb}}^{(k)}$ and (54). For the second term in (53), it follows that

$$\begin{aligned}
&2\sum_{i=1}^{m} \psi_{\mathrm{nb}}'(\gamma_{\mathrm{nb},i}^{(k)})\mathbb{E}\big[\big\langle \delta_{\widehat{x},i} - \delta_{x,i}, \nabla^2\phi_i(x_i^{(k)})(\widehat{x}_i^{(k)} - x_i^{(k)})\big\rangle\big] \\
=&\sum_{i=1}^{m}\left\langle \big(\nabla^2\Phi(x^{(k)})^{-\frac{1}{2}}\nabla_{\widehat{x}}\Psi^{(k)}\big)_i, \nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}\mathbb{E}[\delta_{\widehat{x},i} - \delta_{x,i}]\right\rangle \\
=&\sum_{i=1}^{m}\left\langle \big(\nabla^2\Phi(x^{(k)})^{-\frac{1}{2}}\nabla_{\widehat{x}}\Psi^{(k)}\big)_i, g_{\mathrm{nb},i}^{(k)}\right\rangle = -\beta_{\mathrm{nb}}\big\|\nabla_{\widehat{x}}^{\mathrm{loc}}\Psi_{\mathrm{nb}}^{(k)}\big\|^*_{2,\tau(\widehat{x}^{(k)})+\infty},
\end{aligned}$$

which relies on the fact

$$(\nabla_{\widehat{x}}\Psi(\widehat{x}_i^{(k)}, x_i^{(k)}))_i = 2\psi_{\mathrm{nb}}'(\gamma_{\mathrm{nb},i}^{(k)})\nabla^2\phi_i(x_i^{(k)})(\widehat{x}_i^{(k)} - x_i^{(k)}),$$

the definition of g_{nb} and dual norm. For the third term in (53), we simply reuse (57) and get

$$\begin{aligned}
&\sum_{i=1}^{m} \psi_{\mathrm{nb}}'(\gamma_{\mathrm{nb},i}^{(k)})\mathbb{E}\big[\big\langle \widehat{x}_i^{(k+1)} - x_i^{(k+1)}, (\nabla^2\phi_i(x_i^{(k+1)}) - \nabla^2\phi_i(x_i^{(k)}))(\widehat{x}_i^{(k+1)} - x_i^{(k+1)})\big\rangle\big] \\
\lesssim&(\alpha_{\mathrm{nb}}^2 + \beta_{\mathrm{nb}}^2)\sum_{i=1}^{m} \psi_{\mathrm{nb}}'(\gamma_{\mathrm{nb},i}^{(k)})\mathbb{E}\big[\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}\delta_{x,i}\|\big] \\
\lesssim&\alpha_{\mathrm{nb}}^2\big\|\nabla_\gamma \Psi_{\mathrm{nb}}^{(k)}\big\|^*_{\tau(\widehat{x}^{(k)})+\infty}\big\|\big(\mathbb{E}\big[\|\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}\delta_{x,i}\|\big]\big)_{i\in[m]}\big\|_{\tau(\widehat{x}^{(k)})+\infty} \le \alpha_{\mathrm{nb}}^2\beta\big\|\nabla_\gamma \Psi_{\mathrm{nb}}^{(k)}\big\|^*_{\tau(\widehat{x}^{(k)})+\infty}
\end{aligned}$$

For the quadratic parts in (58), we use the same method and mixed norm bounds to obtain

$$\sum_{i=1}^{m} \psi_{\mathrm{nb}}'(\gamma_{\mathrm{nb},i}^{(k)})\mathbb{E}\left[\delta_{\gamma_{\mathrm{nb},i}}^2\right] \lesssim \alpha_{\mathrm{nb}}^2\beta_{\mathrm{nb}}^2\big\|\nabla_\gamma \Psi_{\mathrm{nb}}^{(k)}\big\|^*_{\tau(\widehat{x}^{(k)})+\infty}.$$

Inserting all the pieces back to (58), it follows that

$$\mathbb{E}\big[\Psi_{\rm nb}^{(k+1)}\big] \le \Psi_{\rm nb}^{(k)} - \frac{\beta_{\rm nb}}{2}\big\|\nabla_{\widehat{x}}^{\rm loc}\Psi_{\rm nb}^{(k)}\big\|_{2,\tau(\widehat{x}^{(k)})+\infty}^{*} + O\big(C^{-1}\alpha_{\rm nb}\beta_{\rm nb} + \lambda_{\rm nb}\alpha_{\rm nb}^2\beta_{\rm nb}^2\big)\big\|\nabla_{\gamma}\Psi_{\rm nb}^{(k)}\big\|_{\tau(\widehat{x}^{(k)})+\infty}^{*}$$

also due to $\lambda_{\rm nb}\big\|\delta_{\gamma_{\rm nb}}\big\|_{\infty} \lesssim \lambda_{\rm nb}\alpha_{\rm nb}\beta_{\rm nb} \le 1$. It still remains to build a norm bridge. For this, because $\big(\nabla_{\widehat{x}}^{\rm loc}\Psi_{\rm nb}^{(k)}\big)_i = 2\nabla^2\phi_i(x_i^{(k)})^{\frac{1}{2}}(\widehat{x}_i^{(k)} - x_i^{(k)})\psi_{\rm nb}'(\gamma_{{\rm nb},i}^{(k)})$, we lower-bound its dual norm

$$\begin{aligned}
\big\|\nabla_{\widehat{x}}^{\rm loc}\Psi_{\rm nb}^{(k)}\big\|_{2,\tau(\widehat{x}^{(k)})+\infty}^{*} =& 2\sum_{i=1}^{m}\big(\nabla_{\gamma}\Psi_{\rm nb}^{(k)}\big)_i\big(\nabla_{\gamma}\Psi_{\rm nb}^{(k)}\big)_i^{\flat(\tau(\widehat{x}^{(k)}))}\sqrt{\gamma_{{\rm nb},i}^{(k)}} \\
\ge& 2\sqrt{\nu}\left(\sum_{i=1}^{m}\big(\nabla_{\gamma}\Psi_{\rm nb}^{(k)}\big)_i\big(\nabla_{\gamma}\Psi_{\rm nb}^{(k)}\big)_i^{\flat(\tau(\widehat{x}^{(k)}))} - \sum_{i:\gamma_{{\rm nb},i}^{(k)}<\nu}\big(\nabla_{\gamma}\Psi_{\rm nb}^{(k)}\big)_i\big(\nabla_{\gamma}\Psi_{\rm nb}^{(k)}\big)_i^{\flat(\tau(\widehat{x}^{(k)}))}\right) \\
\ge& 2\sqrt{\nu}\left(\big\|\nabla_{\gamma}\Psi_{\rm nb}^{(k)}\big\|_{\tau(\widehat{x}^{(k)})+\infty}^{*} - \lambda_{\rm nb}e^{\lambda_{\rm nb}\nu}m\right)
\end{aligned}$$

using a simplification of the method to prove Lemma 35 in the second line. Now we choose $\nu = \alpha_{\rm nb}^2/100$. It holds that $\sqrt{\nu}\beta_{\rm nb}\lambda_{\rm nb}e^{\lambda_{\rm nb}\nu}m \le m^2$. Also, from the definition of the dual norm, we have

$$\big\|\nabla_{\gamma}\Psi_{\rm nb}^{(k)}\big\|_{\tau(\widehat{x}^{(k)})+\infty}^{*} \ge \frac{\big\|\nabla_{\gamma}\Psi_{\rm nb}^{(k)}\big\|_1}{1+C_{\rm norm}\sqrt{2d}} = \frac{\lambda_{\rm nb}}{1+C_{\rm norm}\sqrt{2d}}\Psi_{\rm nb}^{(k)}.$$

Inserting the above computation into the preceding one-step bound and using $\beta_{\rm nb} = C\alpha_{\rm nb}\beta$ yields

$$\begin{aligned}
\mathbb{E}[\Psi_{\rm nb}^{(k+1)}] \le& \Psi_{\rm nb}^{(k)} - \left(\frac{\alpha_{\rm nb}\beta_{\rm nb}}{10} - O\left(C^{-1}\alpha_{\rm nb}\beta_{\rm nb} + \lambda_{\rm nb}\alpha_{\rm nb}^2\beta_{\rm nb}^2\right)\right)\big\|\nabla_{\gamma}\Psi_{\rm nb}^{(k)}\big\|_{\tau(\widehat{x}^{(k)})+\infty}^{*} + m^2 \\
\le& \left(1 - \frac{\theta\beta\log(mT)}{\sqrt{d}}\right)\Psi_{\rm nb}(\widehat{x}^{(k)}, x^{(k)}) + m^2
\end{aligned}$$

□

Proof of Lemma 52. The first property follows from the first two points in Lemma 51, Corollary 27, Lemma 23, Lemma 24, and also Lemma 22. The additional block size factor is absorbed by $\widetilde{O}(1)$. Similar, The rest two properties use Point 3 in Lemma 51, the mentioned lemmas, and also the implementation. □