GRADIENT DESCENT PROGRAM: UNIVERSALITY AND LANDSCAPE

Anonymous authors
Paper under double-blind review

ABSTRACT

How efficiently can gradient descent execute an algorithm, and when does a higher-dimensional parametrization help it optimize? We show that every bounded-fan-in Boolean circuit compiles into a strongly convex objective whose unique minimizer encodes the circuit's gate values. From an initialization encoding a Boolean state, gradient descent reaches this minimizer in a number of exact-arithmetic steps proportional to the depth. Through gradient descent on one fixed strongly convex objective, we show that it is possible to simulate polynomial-space computation. Even though the trajectory converges to a known minimizer, determining the answer encoded at a specified future iteration is PSPACE-complete when the iteration count is given in binary, both in exact arithmetic and for a specified finite-precision implementation. A separate clock-free construction evaluates finite-delay Boolean circuits by projected gradient descent with linear total work using cached derivatives. We then lift Boolean circuits to bounded-trace Gram matrices, connecting execution to optimization over semidefinite relaxations. For these relaxations, we construct fixed losses on which ordinary gradient descent finds approximately feasible solutions or certifies infeasibility in polynomially many exact-arithmetic steps. For any $\epsilon > 0$, sufficient overparameterization gives every point more than ϵ above the global relaxation optimum many directions of negative curvature along which descent can escape. Applications include matrix factorization and completion, sensing, clustering, cuts, constraint satisfaction, moment hierarchies, and neural parameterizations.

1 INTRODUCTION

Gradient methods are widely used in machine learning and continuous optimization, supported by hardware and software that are increasingly organized around backpropagation through neural models (Jouppi et al., 2023; Baydin et al., 2018). Despite this infrastructure, their use in combinatorial optimization remains less direct, because discrete choices obstruct differentiation and a continuous surrogate must still allow feasible solutions of controlled quality to be recovered (Bengio et al., 2021). Promising examples are scattered across routing, satisfiability, integer programming, and graph optimization (Kool et al., 2019; Selsam et al., 2019; Gasse et al., 2019; Khalil et al., 2017).

These examples suggest three complementary directions. First, networks learn solution constructions or search decisions from labeled instances, solver traces, or rewards (Kool et al., 2019; Selsam et al., 2019; Selsam & Bjørner, 2019; Gasse et al., 2019). Second, gradient methods optimize continuous representations, including probability distributions and *relaxations* that enlarge the discrete feasible set (Karalias & Loukas, 2020; Kyrillidis et al., 2021). An important class of relaxations consists of semidefinite programs (SDPs), which optimize a linear objective over positive semidefinite (PSD) matrices subject to linear constraints. Such a matrix is the Gram matrix of a set of vectors and records their inner products. For a maximization problem, the relaxed optimum upper-bounds the discrete optimum, and a rounding procedure converts the matrix solution into a discrete assignment. An approximation guarantee then bounds the returned score relative to the optimum, in expectation when the rounding is randomized (Goemans & Williamson, 1995). SATNet and OptGNN incorporate these relaxations into trainable models (Wang et al., 2019; Yau et al., 2024). Third, neural algorithmic models can learn to reproduce the intermediate steps of known algorithms (Veličković et al., 2020). Together, these directions motivate the following question.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

How efficiently can gradient descent execute an algorithm, and when can richer representations recover the guarantees of continuous relaxations?

This question separates an *execution picture* from a *landscape picture*. In the execution picture, the variables encode working memory and descent implements algorithmic updates, so the trajectory reproduces the computation with controlled memory, work, and precision. In the landscape picture, the variables represent candidate solutions, and the question is whether the objective's geometry excludes poor local optima and supports a guarantee after rounding. Figure 1 contrasts these two roles of gradient descent.

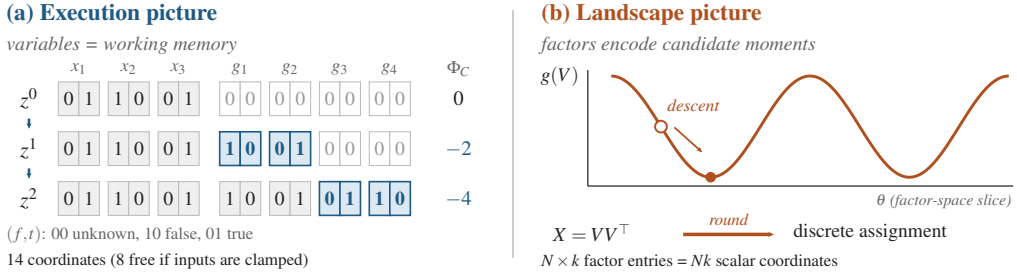


Figure 1: Execution and landscape. (a) Seven wires use 14 false/true rails; two gates resolve per step. (b) The convex matrix loss X_{12}^2 becomes the nonconvex factor slice $\cos^2\theta$, where θ is the angle between two planar unit rows. Both wells yield the same Gram matrix. Rounding is separate.

Our contributions. We address these questions through the following results.

1. We demonstrate the expressive power of gradient descent by compiling Boolean circuits into gradient-descent computations using memory and total work linear in the circuit size, provided unchanged updates are cached (Theorem 2.1). We also give a strongly convex construction in which the unique minimizer encodes the circuit's output. Starting from the prescribed Boolean initialization, gradient descent reaches this minimizer in at most as many exact-arithmetic steps as the depth of the circuit.
2. We extend the compiler from finite circuit evaluation to computations that repeatedly update and reuse their state. This lets a fixed gradient-descent trajectory simulate polynomial-space computation for arbitrarily many steps. As a consequence, even for a strongly convex energy with a known minimizer, predicting a designated output bit after T exact gradient-descent steps is PSPACE-complete when T is given in binary (Theorem 3.1).
3. For smooth semidefinite relaxations, we prescribe a regularized matrix factorization and initialization for which ordinary gradient descent approaches the convex optimum to any prescribed accuracy in exact arithmetic, with a computable certificate (Theorem 4.1). We also show how increasing the factor dimension limits the gap between local minima and the convex optimum (Theorem 4.3).

1.1 Related work

The execution picture builds on computational universality for gradient and neural systems. Abbe & Sandon (2023) establish polynomial-time universality of SGD on specially constructed networks and describe it as P-completeness, and the same line of work identifies precision conditions under which minibatch SGD and full-batch GD can simulate efficient sample-based learning (Abbe et al., 2021). Their analysis crucially depends on careful initialization in a nonconvex landscape. Chatziafratis et al. (2019) study eventual coordinate positivity under online gradient descent. Fearnley & Savani (2018) prove PSPACE-completeness for all-switches strategy improvement through iterated-circuit problems, and Balaji et al. (2019) prove that computing an optimal policy of a finite-horizon Markov decision process is EXPTIME-complete when the horizon is given in binary. Our prediction result adds a fixed strongly convex energy in exact and rounded arithmetic. Dissipative systems, recurrent graph networks, and resistive networks give further comparisons (Šíma & Orponen, 2001; Ahvonen et al., 2024; Rosenbluth & Grohe, 2026; Scellier & Mishra, 2025).

The landscape picture builds on SDP factorization and gradient-based barrier methods (Burer et al., 2002; 2003; Journée et al., 2010). Existing guarantees cover linear elliptope objectives (Mei et al., 2017), smooth SDP manifolds (Boumal et al., 2020), and quantitative factor certificates (Zhang et al., 2023). Global GD convergence is also known for positive definite matrix square roots (Jain et al., 2017) and for regularized Bures–Wasserstein barycenters (Altschuler et al., 2021, Theorem 4, corrected version), whose congruence update lifts to ordinary GD on matrix factors. Polynomial approximations support stochastic spectral gradients (Han et al., 2018). Low-rank recovery relies on assumptions about the measurements or observations (Ge et al., 2017), while factored optimization has local guarantees under additional conditions (Bhojanapalli et al., 2016). Our square-factor construction uses one fixed regularized loss and a scaled-identity initialization to approximate the bounded convex optimum, and its matrix-fitting guarantee concerns objective value rather than recovery. OptGNN and the Lasserre hierarchy supply structured comparisons for constraint satisfaction (Yau et al., 2024; Lasserre, 2001a; Laurent, 2003).

2 Algorithms as gradient descent programs

We first show how gradient descent can carry out a computation by encoding working memory as the state of a descent trajectory. A simulation specifies a differentiable objective $\Phi\colon \mathbb{R}^d \to \mathbb{R}$ on d state coordinates, a product box $K \subseteq \mathbb{R}^d$, an encoded initial state $z^0 \in K$, and a decoder for the final state. We then run ordinary simultaneous projected gradient descent (PGD),

$$z^{t+1} = \Pi_K\big(z^t - \nabla\Phi(z^t)\big), \tag{1}$$

with every derivative evaluated at the old state. The objective measures progress through the encoded computation.

2.1 Making a gradient execute an update

The first obstacle is feedback: in a circuit, a gate reads its inputs and writes its output, leaving the input unchanged. However, gradient descent on, for example, $(y - g(x))^2/2$ generally updates both y and x, so an output mismatch can alter previously computed input values. We suppress this feedback at encoded states with the polynomials

$$q(u) = 3u^2 - 2u^3, \qquad r(u) = u - q(u),$$

which satisfy $q(b) = b$, $q'(b) = 0$, $r(b) = 0$, and $r'(b) = 1$ at $b \in \{0, 1\}$. Thus q reads a bit without passing a derivative through it, while r directs the derivative to the coordinate being written. If R_i is a polynomial update rule for coordinate i, the factor $r(z_i)[q(z_i) - R_i(q(z))]$ has gradient $(b_i - R_i(b))e_i$ at every Boolean state b. Summing these factors makes one unit gradient step execute all rules at once. An additional term $V(q(z))$ tracks the objective's Boolean values as the gates get updated, without disturbing the updates (Lemma A.1).

2.2 Progress from resolved wires

To make the objective measure computational progress, we count the wires whose values have been resolved. Consider a circuit with B input wires and S two-input AND and XOR gates, possibly with a constant-one wire. We represent each wire by two coordinates. One indicating false and the other true. Setting both to zero indicates that the wire's value is unknown. This is called (*dual-rail* encoding),

$$\underbrace{(0,0)}_{\text{unknown}} \longrightarrow \underbrace{(1,0)}_{\text{false}} \quad \text{or} \quad \underbrace{(0,1)}_{\text{true}},$$

with inputs initialized to their given values and internal rails to zero. An AND gate resolves to false as soon as one input is known to be false, whereas an XOR gate waits for both inputs. The resulting rail rules R_i have degree at most four and read at most four parent coordinates. The circuit may contain cycles, provided that repeated simultaneous updates determine every wire's value within D rounds. This *finite-delay* condition holds for acyclic circuits and for the cyclic circuits that simulate bounded-word parallel programs (Heath, 2023). With I indexing the $2S$ internal rails, define

$$\Phi_C(z) = -\sum_{i \in I} q(z_i) + \sum_{i \in I} r(z_i)\big[q(z_i) - R_i(q(z))\big]. \tag{2}$$

At each state reached by the simulation, the second sum is zero. When a gate's value becomes known, one of its two coordinates changes from zero to one, so the objective decreases by one. Each gate's value is determined at most once. The number of known gate values therefore measures progress without an auxiliary clock, as the next theorem makes precise.

Theorem 2.1 (Descent without an additional clock)**.** *For a finite-delay circuit as above, Φ_C is an explicit polynomial of degree at most 15 with $O(B+S)$ coordinates and factors, each involving at most five coordinates. Unit-step PGD on the unit cube from the specified rail initialization reproduces the circuit output within D iterations and then remains fixed, and every reached step satisfies*

$$\Phi_C(z^{t+1}) - \Phi_C(z^t) = -\|z^{t+1} - z^t\|_2^2 = -\|\nabla\Phi_C(z^t)\|_2^2. \tag{3}$$

An implementation that caches unchanged factor derivatives uses $O(B+S)$ total work and storage, including initialization and readout, with words of $O(\log(B+S))$ bits.

Proof sketch. The scalar lift turns unit PGD into the exact rail update. Known information only grows, so each gate changes one rail once. At every reached vertex the correction factors vanish and Φ_C equals minus the number of resolved gates, giving (3). Each factor reads constantly many coordinates and therefore needs constantly many refreshes, giving the work bound (Appendix A, which includes a worked circuit with a feedback edge). Each R_i has degree ≤ 4; q, r have degree 3, so Φ_C has degree ≤ 15.

The work bound depends on caching: scanning every factor in every round would cost $O((B+S)(D+1))$. The descent identity holds along the specified Boolean trajectory, so fractional initializations need a separate analysis. A small change to the value term makes the terminal state a strict local minimum (Proposition A.5).

2.3 Strongly convex constructions

Theorem 2.1 gives universality but stores the computation in the trajectory of a nonconvex objective. To place the circuit's answer at an optimum, we instead use a strongly convex objective with coordinate scales that shrink with depth.

Proposition 2.2 (Circuit values at a strongly convex minimizer)**.** *Consider an acyclic AND/OR/NOT/COPY circuit of total size S_{tot} and depth $D \ge 1$ with $r \ge 1$ sink outputs and at most two distinct parents per gate. For each clamped input there is an explicit energy on $O(S_{\mathrm{tot}})$ coordinates with $O(S_{\mathrm{tot}})$ local factors that is globally $C^{1,1}$, $3/4$-strongly convex, and $5/4$-smooth, so that unit-step GD is globally $1/4$-contractive, and from every scaled Boolean start GD reaches the unique minimizer, which encodes all gate values, within D steps. The local numerical data use $O(D + \log r)$ bits, and dense execution costs $O(S_{\mathrm{tot}} D)$ operations.*

Proof sketch. Gates later in the circuit use progressively smaller powers-of-two scales, limiting their effect on earlier gates. At encoded bit values, the gate functions have zero slope in their inputs, allowing gradient descent to reproduce the intended Boolean updates.. A Hessian bound then gives strong convexity, and induction on depth identifies the minimizer (Appendix B).

For computations that repeatedly rewrite their state, we use a third construction in which the intermediate gradient-descent iterates encode the computation. For a synchronous Boolean map $F\colon \{0,1\}^m \to \{0,1\}^m$ with at most two inputs per register, one fixed $C^{1,1}$ energy on $2m$ coordinates is $3/4$-strongly convex and $5/4$-smooth. Unit-step GD from $c_i^0 = 1$, $x_i^0 = b_i$ has the exact orbit

$$c_i^t = \tau^t, \qquad x_i^t = \tau^t (F^t(b))_i, \tag{4}$$

where $\tau = 2^{-O(\log(2+\Delta))}$ and Δ is the maximum number of factors reading one register (Theorem C.1). The number of coordinates stays fixed as the horizon grows, while the decoding margin $\tau^t/2$ shrinks. Heath's compiler connects these constructions to general programs: it turns a bounded-word program with input-size parameter n, work budget W and parallel time T into a circuit with $S = O(W \log^4 n)$ and $D = O(T \log^3 n)$ (Heath, 2023) (Appendix A.8). The recurrent construction and its rounded version will let us study the difficulty of predicting these computations.

3 PREDICTION COMPLEXITY UNDER CONTRACTION

The strongly convex constructions of Section 2 converge rapidly because their conditioning is fixed. In the recurrent construction, the minimizer is even known in advance, but the computation is encoded in the trajectory approaching it. Rapid convergence does not make that computation easy to predict, even when the recurrent circuit consists of constant-size blocks arranged along a line, each communicating only with its two neighbors. To make this distinction precise, we use the *prediction problem* familiar from cellular automata: given an initial tape over a fixed finite alphabet and an integer $T \geq 0$ in binary, decide whether a designated output bit equals one after T complete recurrent blocks, with fixed modules and initialization rule. The theorem below classifies this problem both in exact arithmetic and for a specified rounded implementation with constant precision and $O(\log(T+2))$ bits per exponent (Appendix D).

Theorem 3.1 (PSPACE-complete prediction with communication along a path)**.** *There is a fixed finite family of ReLU message-passing modules operating on the path $1-2-\cdots-M$ with marked endpoints and left and right ports, such that each vertex stores a constant number of real coordinates and each recurrent block uses a constant number of nearest-neighbor communication stages. For each M, a fixed local rule defines an energy E_M that does not depend on the input tape or on the iteration count. This energy is globally $C^{1,1}$, $3/4$-strongly convex, and $5/4$-smooth, with unique minimizer zero, and its unit-step gradient map is $1/4$-contractive. In exact arithmetic, along trajectories from the prescribed encoded initializations, each recurrent block implements one gradient-descent step on E_M.*

Given an initial tape and an integer $T \geq 0$ in binary, deciding whether the output bit at vertex 1 equals one after T recurrent blocks is PSPACE*-complete under polynomial-time many-one reductions, and the same classification holds for the specified finite-precision rounded implementation.*

Proof sketch. For hardness, we turn a fixed polynomial-space machine into a local tape rule, which the rounded compiler simulates through an exponentially long run. The tape and the binary horizon still have polynomial descriptions. For membership, the Boolean registers and an iteration counter fit in polynomial space (Appendix E).

After T steps, the two possible bit values are separated by only about τ^T, while T itself can be specified using only $O(\log T)$ bits. Therefore, rapid convergence of the trajectory to zero does not make the encoded bit easy to predict. In fact, a polynomial-time algorithm for predicting that bit would imply $\mathrm{P} = \mathrm{PSPACE}$. This shows that a strongly convex objective can be easy to optimize while its intermediate gradient-descent trajectory remains hard to predict.

Two lines of prior work are closely related. Universality results for gradient-based dynamics show that computation can be encoded in optimization trajectories, but use carefully chosen initializations on nonconvex landscapes (Abbe & Sandon, 2023; Abbe et al., 2021). Separately, hardness results for iterated circuits show that predicting a computation at a time T given in binary can be PSPACE-hard; we review these in Section 1. Our result combines these two perspectives: the trajectory of ordinary gradient descent is PSPACE-hard to predict even for a fixed local scalar energy that is uniformly strongly convex.

4 REPRESENTATIONS THAT SUPPORT GRADIENT DESCENT

The execution results start from a supplied algorithm, so the quality of the answer is inherited from that algorithm. To derive a solver from the task itself, we first write down the constraints that the circuit's wire values must satisfy. For a finite circuit, or a recurrent circuit unrolled for a specified number of iterations, these constraints encode the gate relations together with prescribed inputs and outputs (e.g. $z = xy$ for an AND gate with x, y input and z output). We then lift this constraint system to a matrix representation and study the landscape of the resulting optimization problem. After encoding wires by signs, we relax these relations to affine constraints on their moments, which are the expected products of wire values. This gives a PSD matrix called *moment matrix* which has diagonal one because each represented product of signs squares to one.

These moments record correlations that determine what a gradient can see. For intuition, consider the cut problem. On a single edge, independently uniformly sampling labels obtains a cut with probability $1/2$, whereas sampling a uniform label and assigning the opposite label to the other vertex

always cuts the edge. Both sampling rules have zero marginal means, so those means alone do not distinguish their scores. The expected cut score has zero gradient when independently sampled labels have zero means, while a correlation coordinate exposes an improvement to first order (Appendix F).

Vector embeddings, also called *factors*, $V \in \mathbb{R}^{d\times k}$ represent these correlations through a Gram matrix $X = VV^\top \in \mathbb{R}^{d\times d}$. The number of columns k is the *search rank*; the matrix rank may be smaller. Below, we give two theorems that quantify how the factor dimension affects optimization of smooth convex matrix losses, including smoothed affine violations.

For any prescribed accuracy, the first theorem specifies a square factor $V \in \mathbb{R}^{d\times d}$, a fixed regularized loss, and a scaled-identity initialization $V_0 = I_d/\sqrt{d}$ for which ordinary gradient descent reaches that accuracy in finitely many exact-arithmetic steps. The algorithm stops when the gradient norm falls below an explicit threshold guaranteeing that the objective value is at most some ε above the global minimum of the matrix relaxation.

The second gives an explicit bound on the gap to the convex optimum. For a normalized factor with search rank $k \geq 2$, this gap is at most $M/(k-1) + \eta/2$, provided the Hessian along norm-preserving directions has no eigenvalue below $-\eta$. Here M bounds the curvature of the convex matrix loss. Increasing the number of columns therefore reduces the first term, while the second measures the remaining negative curvature.

We then recover and extend many known results as corollaries of these two theorems (Table 1). For Boolean tasks, the size and tightness of the relaxation, together with rounding, determine the resulting guarantee.

4.1 Constant-step gradient descent on a square factor

Throughout, $\mathcal{D}_d$ denotes the set of *density matrices*, which are $d \times d$ PSD matrices of unit trace.

We first use a square factor of rank d to optimize a smooth convex objective over $\mathcal{D}_d$ to prescribed accuracy using an explicit stopping rule based on the gradient norm. Let ϕ be convex and C^2 near $\mathcal{D}_d$ with $\|\nabla\phi\|_{\mathrm{op}} \leq L$ and $0 \leq \nabla^2\phi[H,H] \leq M\|H\|_F^2$ there. For a nonzero square factor $V \in \mathbb{R}^{d\times d}$, we normalize the Gram matrix to impose unit trace. To keep the iterates uniformly positive definite, we add a regularizer with weight $\mu > 0$ and degree p, obtained by truncating the series of the barrier $-\mu \log\det X$:

$$X(V) = \frac{VV^\top}{\|V\|_F^2}, \qquad \mathcal{L}_p(V) = \phi(X(V)) + \mu \sum_{j=1}^{p} \frac{\mathrm{tr}[(I_d - X(V))^j]}{j}. \tag{5}$$

Log-determinant barriers have been used in gradient-based methods for semidefinite optimization (Burer et al., 2002; 2003). Here we use a finite polynomial truncation of this barrier, yielding a regularizer that is polynomial in X, remains finite even at singular matrices, and has derivatives that can be evaluated using matrix products. The theorem shows that this finite regularizer preserves a uniform lower bound on the eigenvalues along the entire run.

Theorem 4.1 (Ordinary GD for convex objectives on density matrices)**.** *Set $\Lambda = L + \mu d$, $r = \mu/(4\Lambda)$, and $p = \lceil r^{-1}\log(64\Lambda/\mu)\rceil$, and let $M_F \geq 0$ bound the Hessian of $U \mapsto \phi(UU^\top)$ by $2L + M_F$ at unit factors with $UU^\top \succeq rI_d$ (one may take $M_F = 4M$). Run ordinary GD as:*

$$V_{t+1} = V_t - \eta\nabla\mathcal{L}_p(V_t), \qquad V_0 = I_d/\sqrt{d}, \qquad 0 < \eta \leq [64\Lambda + M_F]^{-1}. \tag{6}$$

Then every iterate satisfies $X(V_t) \succeq rI_d$ and $1 \leq \|V_t\|_F^2 \leq 1 + 4\eta L$; the regularized objective has a unique minimizer $X_* \in \mathcal{D}_d$ and $\mathcal{L}_p(V_t)$ converges to its value at the geometric rate $2L(1 - 4\mu\eta/(1 + 4\eta L))^t$; and with $G_t = \|V_t\|_F\nabla\mathcal{L}_p(V_t)$, every iterate certifies

$$\phi(X(V_t)) - \min_{\mathcal{D}_d}\phi \leq \mu d + \frac{\|G_t\|_F^2}{8\mu}. \tag{7}$$

Proof sketch. The polynomial preserves a lower bound on the smallest eigenvalue along every trial step, so the loss is smooth where the iterates live. In this region, a weighted log-determinant inequality bounds the loss gap by the squared gradient, giving geometric convergence and the gradient certificate. Scale invariance controls the growth of $\|V_t\|_F$, so the step-size coefficient $\eta/\|V_t\|_F^2$ for the normalized factor stays bounded away from 0. □

The proof is in Appendix G. To use the certificate, choose $\mu = \varepsilon/(2d)$ and stop when $\|G_t\|_F^2 \le 4\mu\varepsilon$. This occurs within $\widetilde{O}(d(L+M)/\varepsilon)$ steps and leaves ϕ within ε of its minimum. Apart from evaluating $\nabla\phi$, each gradient requires $O(\log p)$ matrix products and no matrix solves or spectral subroutines.

In perspective, gradient-based SDP barriers already read feasibility and complementarity from transformed gradients (Burer et al., 2003, Corollary 1), with global convergence through a sequence of barrier subproblems (Burer et al., 2003, Algorithm GLB and Theorem 5). Factored GD already has finite convergence bounds under a suitable initialization and matching search rank (Bhojanapalli et al., 2016, Theorem 5 and Assumption (A1)). The contribution here is the joint prescription of a fixed polynomial regularizer, scaled-identity initialization, constant step, uniform lower bound on the eigenvalues, and finite certificate for smooth convex objectives on $\mathcal{D}_d$. In comparison, the certificate of Zhang et al. (2023, Proposition 10) also combines gradient, Hessian, and rank-deficiency terms; while (7) uses only the regularized gradient along the full-rank trajectory. Projection-free solvers already provide finite accuracy certificates using an eigenvector oracle at every step (Hazan, 2008; Jaggi, 2013), so this construction makes no claim to a faster solver.

Example: semidefinite feasibility. We use bounded semidefinite feasibility as a running example for both theorems. The goal is to find a positive semidefinite matrix $Z \in \mathbb{R}^{n\times n}$ satisfying given affine constraints and $\operatorname{tr} Z \le R$; or to certify that no such matrix exists. For $R > 0$, we write $Z = RX_{1:n,1:n}$, where $X_{1:n,1:n}$ is the submatrix of the first n rows and columns of a density matrix $X \in \mathcal{D}_d$ with $d = n + 1$. This representation enforces positive semidefiniteness and the trace bound, with the extra diagonal entry absorbing the unused trace.

Write the remaining constraints as $a_i(X) = b_i + \langle B_i, X\rangle \le 0$, $1 \le i \le m$. A positive residual $a_i(X)$ measures how much constraint i is violated. Introducing $a_0(X) = 0$, we define the largest violation and its smallest achievable value by

$$v(X) = \max_{0\le i\le m} a_i(X), \qquad \gamma_* = \min_{X\in\mathcal{D}_d} v(X).$$

Thus $v(X) = 0$ exactly when X satisfies all constraints, while $\gamma_* > 0$ certifies infeasibility.

To apply the smooth optimization theorem, we replace the nonsmooth maximum by

$$\phi_\tau(X) = \tau \log \sum_{i=0}^{m} e^{a_i(X)/\tau}, \qquad v(X) \le \phi_\tau(X) \le v(X) + \tau \log q,$$

where $q = m + 1$. The following corollary bounds γ_* from both sides: the current matrix gives an upper bound, while the gradient certificate gives a lower bound. Set $L = \max_{1\le i\le m} \|B_i\|_{\mathrm{op}} > 0$.

Corollary 4.2 (Bounded semidefinite feasibility)**.** *For $0 < \varepsilon \le L$, apply Theorem 4.1 to $\phi = \phi_\tau$ with*

$$\tau = \frac{\varepsilon}{4\log q}, \qquad \mu = \frac{\varepsilon}{4d}, \qquad \eta = [64\Lambda + 4L^2/\tau]^{-1}, \quad \Lambda = L + \mu d.$$

At each iterate, write $X = X(V)$, $G = \|V\|_F \nabla\mathcal{L}_p(V)$, and $\pi_i = e^{a_i(X)/\tau}/\sum_{j=0}^m e^{a_j(X)/\tau}$. Then

$$\underline{\gamma}(V) := \sum_{i=0}^{m} \pi_i a_i(X) - \mu d - \frac{\|G\|_F^2}{8\mu} \le \gamma_* \le v(X).$$

After $\widetilde{O}(d(L/\varepsilon)^2 \log q)$ exact-arithmetic gradient steps, these bounds differ by at most ε. If $v(X) \le \varepsilon$, the returned matrix violates no constraint by more than ε. Otherwise, $\underline{\gamma}(V) > 0$ certifies that no matrix within the trace bound satisfies all constraints.

The same theorem also gives guarantees for matrix fitting under a nuclear-norm constraint (Mazumder et al., 2010), using the standard reformulation over density matrices (Jaggi & Sulovský, 2010); see Corollary H.2. Appendix H gives the iteration bound and specializes the construction to positive semidefinite sensing and matrix completion with exact observed entries. Appendix I treats the Peng–Wei clustering relaxation. For the matrix-fitting problems, recovering a planted low-rank matrix additionally requires assumptions on the measurements (Ge et al., 2017).

Table 1: Selected applications of Theorems 4.1 (first row) and 4.3(b) (last two); full list and comparisons in Tables 2–3. **A**: new prescribed-GD guarantee; **E**: new quantitative extension. The landscape gap is at most $(R + C/2)/(k-1) + N\zeta/2$ when $\operatorname{Hess} F \preceq \zeta I$; the listed constants are $R + C/2$. Here ρ is the feasibility penalty and $\tau > 0$ the contrastive temperature.

Problem and precedent	Coordinates	Guarantee	Status / result
Matrix completion (Candès & Recht, 2009)	$(a+b+1)^2$	observed entries exact, nuclear norm within ε of the minimum via norm-bound feasibility tests and entry repair	**A** Cor. H.7
Lovász theta (Lovász, 1979)	$(n+1)k$	$1 + 21\rho/4$: certified interval for $\vartheta(G)/n$ after an explicit repair	**E** Prop. L.7
Contrastive embeddings (Khosla et al., 2020; Behnia & Thrampoulidis, 2024; Vock et al., 2026)	Nk	$4/\tau + 2/\tau^2$: in-sample loss within the bound of the convex optimum	**E** App. L

4.2 What smaller ranks guarantee

The square factor gives a specified descent algorithm. Smaller factors use fewer coordinates, but can have stationary points with poor objective values. Our second theorem bounds this suboptimality in two representations. The first fixes the trace, while the second fixes every diagonal entry. For the second form, define the elliptope $\mathcal{E}_N = \{X \succeq 0 : X_{ii} = 1\}, \mathcal{M}_{N,k} = (\mathbb{S}^{k-1})^N$, and let $R(A) = \max_{Z \in \mathcal{E}_N} \langle A, Z\rangle - \min_{Z \in \mathcal{E}_N} \langle A, Z\rangle$.

Theorem 4.3 (Landscape bounds at fixed rank)**.** *Let $k \geq 2$ and $\eta \geq 0$. All gradients and Hessians below are Riemannian, with the metric induced by the Frobenius inner product, except for the matrix derivatives of f.*

(a) Unit trace. *Let f be convex and C^2 near $\mathcal{D}_d$, let $W \in \mathbb{R}^{d\times k}$ satisfy $\|W\|_F = 1$, and set $X = WW^\top$, $A = \nabla f(X)$, and $g(W) = f(WW^\top)$. If $\operatorname{Hess} g(W) \succeq -\eta I$ and $\nabla^2 f(X)[H,H] \leq M\|H\|_F^2$ for every symmetric H, where $M \geq 0$, then*

$$f(X) - \min_{\mathcal{D}_d} f \leq \Delta(X) := \langle A, X\rangle - \lambda_{\min}(A) \leq \frac{M}{k-1} + \frac{\eta}{2}. \tag{8}$$

Here Δ is the Frank–Wolfe gap (Jaggi, 2013). If also $\operatorname{grad} g(W) = 0$, the denominator improves to k. The resulting M/k coefficient is sharp for Δ.

(b) Unit diagonal. *Let f be concave and C^2 near $\mathcal{E}_N$, let $V \in \mathcal{M}_{N,k}$ have rows $v_i^\top$, and set $F(V) = f(VV^\top)$ and $A = \nabla f(VV^\top)$. Suppose $\operatorname{Hess} F(V) \preceq \eta I$. For a comparator $X_* = YY^\top \in \mathcal{E}_N$, write the rows of $Y \in \mathbb{R}^{N\times s}$ as $y_i^\top$, and draw $G \in \mathbb{R}^{k\times s}$ with independent $\mathcal{N}(0, 1/k)$ entries. Let U have rows $u_i^\top$, where $u_i = (I - v_i v_i^\top)Gy_i$, and put $H = VU^\top + UV^\top$. If $C \geq 0$ satisfies the curvature condition*

$$-k\,\mathbb{E}\,\nabla^2 f(VV^\top)[H,H] \leq C, \tag{9}$$

where the expectation is over G, then

$$f(X_*) - F(V) \leq \frac{R(A) + C/2}{k-1} + \frac{N\eta}{2}. \tag{10}$$

Taking X_* to be a maximizer bounds the gap to the relaxation optimum. Neither bound requires first-order stationarity.

Proof sketch. For part (a), take a unit bottom eigenvector u of A and a unit vector $z \perp W^\top u$. The tangent direction $uz^\top$ has curvature $-2\Delta(X) + \nabla^2 f(X)[H,H]$, where $H = Wzu^\top + uz^\top W^\top$ and $\|H\|_F^2 = 2\|Wz\|^2$. Averaging over an orthonormal basis of $(W^\top u)^\perp$ gives the stated bound (Appendix J). Part (b) averages over Gaussian directions projected onto the tangent space of each row sphere (Appendix K). □

Example: semidefinite feasibility at smaller search ranks. We return to the semidefinite feasibility problem of Corollary 4.2 and show that with high enough rank, every local minimum has a largest constraint violation within ϵ of the best possible value over the relaxation.

Corollary 4.4 (Rank supplies many negative-curvature directions)**.** *Using notation from Corollary 4.2, for the smoothed feasibility objective ϕ_τ with $\tau = \varepsilon/(4 \log q)$ and $k - 1 \geq 32L^2 \log q/\varepsilon^2$, every factor W either has $\Delta(WW^\top) \leq 3\varepsilon/4$, in which case $\max_i a_i(WW^\top) - \gamma_* \leq \varepsilon$, or the sphere Hessian of $\phi_\tau(WW^\top)$ has at least $\lceil (k-1)/2 \rceil$ eigenvalues below $-\varepsilon/2$.*

Thus, at any point with largest constraint violation exceeding its minimum γ_* over the relaxation by more than ε, the number of negative-curvature directions grows linearly with the rank. Reaching an approximate second-order point with Hessian tolerance below $\varepsilon/2$ therefore guarantees that every constraint violation is at most $\gamma_* + \varepsilon$. For a feasible instance, $\gamma_* = 0$. Gradient descent with small random perturbations can reach such points (Jin et al., 2017; Criscitiello & Boumal, 2019). More escape directions help it leave saddles faster locally (Appendix J), but whether this improves the overall convergence rate remains open.

The new guarantees are the prescribed GD construction (Theorem 4.1), the trace-sphere certificate and negative-curvature count (Theorem 4.3(a) and Corollary 4.4), and the nonlinear unit-row bound (Theorem 4.3(b)).

4.3 Transfer to neural parameterizations

The landscape bound transfers to a neural encoder when its output head can realize every tangent direction of the factor. For the concave function f in Theorem 4.3(b), write $F(V) = f(VV^\top)$ and let $\Psi = [\psi_1, \ldots, \psi_N] \in \mathbb{R}^{h \times N}$ collect the encoder features and let B be a linear head. The rows of V are the normalized head outputs $v_i = z_i/\|z_i\|$ with $z_i = B\psi_i$.

Corollary 4.5 (Normalized neural head)**.** *Suppose $\sigma_N(\Psi) > 0$ and all $z_i \neq 0$. If the Hessian of the head loss satisfies $\nabla_B^2[-F] \succeq -\gamma I$ with the features held fixed and $\gamma \geq 0$, then the bound in Theorem 4.3(b) holds with $\eta = \gamma(\max_i \|z_i\|/\sigma_N(\Psi))^2$, and an exact second-order stationary parameter point maps to an exact second-order stationary factor point.*

The proof is in Appendix M. The appendix also recovers the known prediction-loss certificates for full-rank feature matrices and regularized homogeneous networks (Nguyen & Hein, 2017; Haeffele & Vidal, 2017).

5 Discussion

We asked how efficiently gradient descent can execute an algorithm and when a richer representation recovers the guarantees of a relaxation. On the first question, we gave three constructions. The objective of Theorem 2.1 executes a circuit and measures progress by resolved wires. The static strongly convex compiler stores the answer at its minimizer, while the recurrent construction carries it in a transient whose decoding margin shrinks with time. A trajectory can therefore approach a known minimizer at a fixed geometric rate while its shrinking coordinates encode a PSPACE-hard answer, making the numerical representation and decoder part of the guarantee.

On the second question, additional state supports both an algorithmic guarantee and a landscape bound. With a square factor, the polynomial regularizer keeps the smallest eigenvalue bounded below and gives ordinary constant-step GD a finite stopping guarantee from one scaled-identity start on every instance. This guarantee charges the trace radius, coefficient scale, and matrix products, and certifies feasibility to additive tolerance. At smaller search ranks, extra columns provide many negative-curvature directions, although converting their number into an improved global rate remains open. Transferring this geometry to neural parameterizations requires the encoder to expose the tangent directions used by the factor argument. The transfer corollary therefore identifies the conditioning of the feature matrix as the quantity to monitor during training.

AI USE STATEMENT

In this work, we used OpenAI Codex and Anthropic Claude under our guidance to help develop theoretical models and conceptual frameworks, formulate mathematical claims, propose and refine hypotheses, provide critical ingredients for proving mathematical claims, assist in writing proofs, and give feedback on research methodology. We also used these tools to implement methods and internal computational checks, generate synthetic test data for those checks, and interpret their outputs. These uses concern the development and verification of theoretical results; the paper does not report an experimental study. Translation, dataset cleaning and reformatting, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used these tools to draft manuscript passages, improve readability, identify relevant literature, and format references. We have reviewed all AI-assisted work, including the text, mathematical claims and proofs, references, and computational checks. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendices A–E specify the circuit encodings, update rules, arithmetic models, and proofs for execution and prediction complexity. Appendix G gives the regularizer, initialization, step size, stopping certificate, and arithmetic cost for prescribed gradient descent, with matrix-fitting reductions in Appendix H. Appendices J–M provide the landscape, application, rounding, and neural-transfer proofs and their assumptions.

REFERENCES

Emmanuel Abbe and Colin Sandon. Polynomial-time universality and limitations of deep learning. *Communications on Pure and Applied Mathematics*, 76(11):3493–3549, 2023. doi: 10.1002/cpa.22121. URL https://doi.org/10.1002/cpa.22121.

Emmanuel Abbe, Pritish Kamath, Eran Malach, Colin Sandon, and Nathan Srebro. On the power of differentiable learning versus PAC and SQ learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24340–24351, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cc225865b743ecc91c4743259813f604-Abstract.html.

P.-A. Absil, Robert Mahony, and Ben Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005. doi: 10.1137/040605266. URL https://sites.uclouvain.be/absil/Publi/Convergence_SIOPT/13-60526.pdf.

Veeti Ahvonen, Damian Heiman, and Antti Kuusisto. Descriptive complexity for neural networks via Boolean networks. In *32nd EACSL Annual Conference on Computer Science Logic*, volume 288 of *LIPIcs*, pp. 9:1–9:22, 2024. doi: 10.4230/LIPIcs.CSL.2024.9.

Jason M. Altschuler, Sinho Chewi, Patrik Gerber, and Austin J. Stromme. Averaging on the Bures–Wasserstein manifold: Dimension-free convergence of gradient descent. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22132–22145, 2021. URL https://arxiv.org/abs/2106.08502v3. Theorem statements follow the corrected author version, arXiv:2106.08502v3 (31 October 2023).

Nikhil Balaji, Stefan Kiefer, Petr Novotný, Guillermo A. Pérez, and Mahsa Shirmohammadi. On the complexity of value iteration. In *46th International Colloquium on Automata, Languages, and Programming*, volume 132 of *Leibniz International Proceedings in Informatics*, pp. 102:1–102:15, 2019. doi: 10.4230/LIPIcs.ICALP.2019.102. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICALP.2019.102.

Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018. URL https://jmlr.org/papers/v18/17-468.html.

Tina Behnia and Christos Thrampoulidis. Supervised contrastive representation learning: Landscape analysis with unconstrained features. In *2024 IEEE International Symposium on Information Theory (ISIT)*, 2024. URL https://arxiv.org/abs/2402.18884.

Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: A methodological tour d'horizon. *European Journal of Operational Research*, 290(2):405–421, 2021. doi: 10.1016/j.ejor.2020.07.063.

Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 530–582, 2016. URL https://proceedings.mlr.press/v49/bhojanapalli16.html.

Nicolas Boumal, Vladislav Voroninski, and Afonso S. Bandeira. Deterministic guarantees for Burer–Monteiro factorizations of smooth semidefinite programs. *Communications on Pure and Applied Mathematics*, 73(3):581–608, 2020. doi: 10.1002/cpa.21830. URL https://arxiv.org/abs/1804.02008.

Samuel Burer and Renato D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95:329–357, 2003. doi: 10.1007/s10107-002-0352-8.

Samuel Burer, Renato D. C. Monteiro, and Yin Zhang. Solving a class of semidefinite programs via nonlinear programming. *Mathematical Programming*, 93(1):97–122, 2002. doi: 10.1007/s10107 0100279. URL https://optimization-online.org/wp-content/uploads/2001/10/382.pdf.

Samuel Burer, Renato D. C. Monteiro, and Yin Zhang. A computational study of a gradient-based log-barrier algorithm for a class of large-scale SDPs. *Mathematical Programming*, 95(2):359–379, 2003. doi: 10.1007/s10107-002-0353-7. URL https://sburer.github.io/papers/008-sdpvnlp4.pdf.

Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010. doi: 10.1137/080738 970. URL https://www.math.hkust.edu.hk/~jfcai/paper/SVT.pdf.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009. doi: 10.1007/s10208-009-9045-5. URL https://link.springer.com/article/10.1007/s10208-009-9045-5.

Vaggos Chatziafratis, Tim Roughgarden, and Joshua R. Wang. On the computational power of online gradient descent. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 624–662, 2019. URL https://proceedings.mlr.press/v99/chatziafratis19a.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Christopher Criscitiello and Nicolas Boumal. Efficiently escaping saddle points on manifolds. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/7486cef2522ee03547cfb970a404a874-Abstract.html.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: layers are automatically balanced. In *Advances in Neural Information Processing Systems*, 2018.

Maryam Fazel, Haitham Hindi, and Stephen Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the American Control Conference*, volume 6, pp. 4734–4739, 2001. doi: 10.1109/ACC.2001.945730. URL https://web.stanford.edu/~boyd/papers/pdf/rank_min_heur_sys_approx.pdf.

John Fearnley and Rahul Savani. The complexity of all-switches strategy improvement. *Logical Methods in Computer Science*, 14(4):1–57, 2018. doi: 10.23638/LMCS-14(4:9)2018. URL https://lmcs.episciences.org/4940. Article 9.

Elisabeth Gaar and Franz Rendl. A computational study of exact subgraph based SDP bounds for Max-Cut, stable set and coloring. *Mathematical Programming*, 183:283–308, 2020. doi: 10.1007/s10107-020-01512-2.

Maxime Gasse, Didier Chételat, Nicola Ferroni, Laurent Charlin, and Andrea Lodi. Exact combinatorial optimization with graph convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/d14c2267d848abeb81fd590f371d39bd-Abstract.html.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1233–1242, 2017. URL https://proceedings.mlr.press/v70/ge17a.html.

Michel X. Goemans and David P. Williamson. New 3/4-approximation algorithms for the maximum satisfiability problem. *SIAM Journal on Discrete Mathematics*, 7(4):656–666, 1994. doi: 10.1137/S0895480192243516. URL https://math.mit.edu/~goemans/PAPERS/GoemansWilliamson-1994-NewApproximationAlgorithmsForTheMaximumSatisfiabilityProblem.pdf.

Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995. doi: 10.1145/227683.227684.

Igor Grubišić and Raoul Pietersz. Efficient rank reduction of correlation matrices. *Linear Algebra and its Applications*, 422(2–3):629–653, 2007. doi: 10.1016/j.laa.2006.11.024. URL https://arxiv.org/abs/cond-mat/0403477.

Benjamin D. Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond, 2015. URL https://arxiv.org/abs/1506.07540.

Benjamin D. Haeffele and René Vidal. Global optimality in neural network training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7331–7339, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Haeffele_Global_Optimality_in_CVPR_2017_paper.html.

Insu Han, Haim Avron, and Jinwoo Shin. Stochastic Chebyshev gradient descent for spectral optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/898aef0932f6aaecda27aba8e9903991-Abstract.html.

Elad Hazan. Sparse approximate solutions to semidefinite programs. In *LATIN 2008: Theoretical Informatics*, volume 4957 of *Lecture Notes in Computer Science*, pp. 306–316, 2008. doi: 10.1007/978-3-540-78773-0_27. URL https://doi.org/10.1007/978-3-540-78773-0_27.

David Heath. Parallel RAM from cyclic circuits. *arXiv preprint arXiv:2309.05133*, 2023. URL https://arxiv.org/abs/2309.05133.

Nicholas J. Higham. Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, 22(3):329–343, 2002a. doi: 10.1093/imanum/22.3.329. URL https://eprints.maths.manchester.ac.uk/232/1/paper3.pdf.

Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2002b. doi: 10.1137/1.9780898718027. URL https://doi.org/10.1137/1.9780898718027.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28(1) of *Proceedings of Machine Learning Research*, pp. 427–435. PMLR, 2013. URL https://proceedings.mlr.press/v28/jaggi13.html.

Martin Jaggi and Marek Sulovský. A simple algorithm for nuclear norm regularized problems. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 471–478, 2010. URL https://icml.cc/Conferences/2010/papers/196.pdf.

Prateek Jain, Chi Jin, Sham Kakade, and Praneeth Netrapalli. Global convergence of non-convex gradient descent for computing matrix squareroot. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 479–488, 2017. URL https://proceedings.mlr.press/v54/jain17a.html.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1724–1732. PMLR, 2017. URL https://proceedings.mlr.press/v70/jin17a.html.

Norman P. Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, Cliff Young, Xiang Zhou, Zongwei Zhou, and David Patterson. TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pp. 1–14, 2023. doi: 10.1145/3579371.3589350. URL https://doi.org/10.1145/3579371.3589350.

Michel Journée, Francis Bach, P.-A. Absil, and Rodolphe Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010. doi: 10.1137/080731359. URL https://sites.uclouvain.be/absil/Publi/2008-127_low-rank/73135.pdf.

Nikolaos Karalias and Andreas Loukas. Erdős goes neural: An unsupervised learning framework for combinatorial optimization on graphs. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/49f85a9ed090b20c8bed85a5923c669f-Abstract.html.

Elias Khalil, Hanjun Dai, Yuyu Zhang, Bistra Dilkina, and Le Song. Learning combinatorial optimization algorithms over graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/file/d9896106ca98d3d05b8cbdf4fd8b13a1-Paper.pdf.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.neurips.cc/paper_files/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Wouter Kool, Herke van Hoof, and Max Welling. Attention, learn to solve routing problems! In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByxBFsRqYm.

Anastasios Kyrillidis, Anshumali Shrivastava, Moshe Y. Vardi, and Zhiwei Zhang. Solving hybrid Boolean constraints in continuous space via multilinear Fourier expansions. *Artificial Intelligence*, 299:103559, 2021. doi: 10.1016/j.artint.2021.103559.

Jean B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001a. doi: 10.1137/S1052623400366802.

Jean B. Lasserre. An explicit exact SDP relaxation for nonlinear 0–1 programs. In Karen Aardal and Bert Gerards (eds.), *Integer Programming and Combinatorial Optimization*, volume 2081 of *Lecture Notes in Computer Science*, pp. 293–303. Springer, 2001b. doi: 10.1007/3-540-45535-3_23. URL https://doi.org/10.1007/3-540-45535-3_23.

Monique Laurent. A comparison of the sherali–adams, lovász–schrijver, and lasserre relaxations for 0-1 programming. *Mathematics of Operations Research*, 28(3):470–496, 2003. doi: 10.1287/moor.28.3.470.16391.

Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176:311–337, 2019. doi: 10.1007/s10107-019-01374-3. URL https://link.springer.com/article/10.1007/s10107-019-01374-3.

Eitan Levin, Joe Kileel, and Nicolas Boumal. The effect of smooth parametrizations on nonconvex optimization landscapes. *Mathematical Programming*, 209:63–111, 2025. doi: 10.1007/s10107-024-02058-3. URL https://arxiv.org/abs/2207.03512. First published online in 2024.

László Lovász. On the Shannon capacity of a graph. *IEEE Transactions on Information Theory*, 25 (1):1–7, 1979. doi: 10.1109/TIT.1979.1055985. URL https://web.vu.lt/mif/s.jukna/boolean/Lovasz-theta.pdf.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010. URL https://www.jmlr.org/papers/v11/mazumder10a.html.

Song Mei, Theodor Misiakiewicz, Andrea Montanari, and Roberto Imbuzeiro Oliveira. Solving SDPs for synchronization and MaxCut problems via the Grothendieck inequality. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1476–1515, 2017. URL https://proceedings.mlr.press/v65/mei17a.html.

Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2603–2612. PMLR, 2017. URL https://proceedings.mlr.press/v70/nguyen17a.html.

Jiming Peng and Yu Wei. Approximating K-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18(1):186–205, 2007. doi: 10.1137/050641983. URL https://optimization-online.org/wp-content/uploads/2005/04/1114.pdf.

Eran Rosenbluth and Martin Grohe. Repetition makes perfect: Recurrent graph neural networks match message passing limit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40(30), pp. 25168–25175, 2026. doi: 10.1609/aaai.v40i30.39707. URL https://ojs.aaai.org/index.php/AAAI/article/view/39707.

Benjamin Scellier and Siddhartha Mishra. Universal approximation theorem for nonlinear resistive networks. *Physical Review Applied*, 23(4):044009, 2025. doi: 10.1103/PhysRevApplied.23.044009. URL https://arxiv.org/abs/2312.15063.

Daniel Selsam and Nikolaj Bjørner. Guiding high-performance SAT solvers with unsat-core predictions. In *Theory and Applications of Satisfiability Testing*, volume 11628 of *Lecture Notes in Computer Science*, pp. 336–353, 2019. doi: 10.1007/978-3-030-24258-9_24. URL https://doi.org/10.1007/978-3-030-24258-9_24.

Daniel Selsam, Matthew Lamm, Benedikt Bünz, Percy Liang, Leonardo de Moura, and David L. Dill. Learning a SAT solver from single-bit supervision. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1802.03685.

Jiří Šíma and Pekka Orponen. Computing with continuous-time Liapunov systems. In *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing*, pp. 722–731, 2001. URL https://users.ics.aalto.fi/orponen/papers/stoc_2001.pdf.

Larry J. Stockmeyer and Albert R. Meyer. Word problems requiring exponential time: Preliminary report. In *Proceedings of the Fifth Annual ACM Symposium on Theory of Computing*, pp. 1–9, 1973. doi: 10.1145/800125.804029.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996. doi: 10.1137/1038003. URL https://web.stanford.edu/~boyd/papers/pdf/semidef_prog.pdf.

Petar Veličković, Rex Ying, Matilde Padovano, Raia Hadsell, and Charles Blundell. Neural execution of graph algorithms. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkgKO0EtvS.

Raphael Vock, Edouard Duchesnay, and Benoit Dufumier. A unified geometric framework for weighted contrastive learning, 2026. URL https://arxiv.org/abs/2605.13943.

Po-Wei Wang, Priya Donti, Bryan Wilder, and Zico Kolter. SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6545–6554, 2019. URL https://proceedings.mlr.press/v97/wang19e.html.

Morris Yau, Nikolaos Karalias, Eric Lu, Jessica Xu, and Stefanie Jegelka. Are graph neural networks optimal approximation algorithms? In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2328. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/85db52cc08c5e00cfb1d216b1c85ba35-Abstract-Conference.html.

Gavin Zhang, Salar Fattahi, and Richard Y. Zhang. Preconditioned gradient descent for overparameterized nonconvex Burer–Monteiro factorization with global optimality certification. *Journal of Machine Learning Research*, 24(163):1–55, 2023. URL https://jmlr.org/papers/volume24/22-0882/22-0882.pdf.

APPENDICES

The appendices follow the main text from circuit execution and prediction complexity to matrix optimization, landscape bounds, and neural parameterizations. The feasibility example is developed in both the GD and landscape proofs; the unit-diagonal applications then distinguish penalty bounds, rounding guarantees, and uniform constraint control.

APPENDIX CONTENTS

Circuit execution and prediction complexity **18**

A Proofs for the simulation construction **18**
A.1 The scalar lift . . . 18
A.2 Rails and finite-delay circuits . . . 18
A.3 Exact simulation and decrease . . . 19
A.4 Counting the cost of a full simultaneous gradient . . . 20
A.5 A strict local minimum at the terminal state . . . 20
A.6 Worked example with a feedback edge . . . 21
A.7 Message passing . . . 21
A.8 Computational models and resource bounds . . . 22

B Circuit evaluation at a strongly convex minimizer **23**
B.1 Circuit normalization and coordinate scales . . . 23
B.2 The local energy . . . 24
B.3 Global conditioning . . . 25

C Recurrent computation in a fixed convex energy **27**
C.1 The recurrent model and the local-clock theorem . . . 27
C.2 Gate interpolation and the clock factors . . . 28
C.3 Global regularity and the exact descent orbit . . . 29

D Local rounded gradient computation **31**
D.1 The local factors and their purpose . . . 32
D.2 Global conditioning, including zero clocks . . . 33
D.3 The gradient on Boolean tubes and the numerical invariant . . . 35
D.4 ReLU messages and finite exponent storage . . . 36
D.5 Descent and completion of the local theorem . . . 39

E The prediction problem and its complexity **39**
E.1 Fixed architecture and arithmetic . . . 40
E.2 Exact and rounded evaluation . . . 40
E.3 Hardness and membership . . . 41

Matrix representations and prescribed gradient descent **43**

F Why correlation coordinates matter **43**

G Ordinary gradient descent with a polynomial spectral regularizer **44**
G.1 The polynomial loss and the general theorem . . . 45
G.2 A lower bound on the smallest eigenvalue along every trial step . . . 46
G.3 Descent and a weighted gradient inequality . . . 47
G.4 Affine feasibility and a finite stopping time . . . 50
G.5 Arithmetic implementation . . . 52
G.6 Initialization and search-rank obstructions . . . 52

H Matrix factorization and completion **54**
H.1 A common accuracy prescription . . . 54
H.2 Rectangular factorization over a nuclear-norm ball . . . 55
H.3 Positive semidefinite matrix sensing . . . 57
H.4 Exact observed entries and approximate minimum nuclear norm . . . 58

I The exact log-determinant comparison **60**
I.1 A general convex objective . . . 60
I.2 Affine constraints and a gradient-based certificate . . . 63
I.3 Finite-time accuracy and arithmetic scope . . . 65
I.4 A local conditioning limit . . . 66
I.5 Applications with an explicit trace bound . . . 67

Landscape bounds and applications **69**

J Certificates and negative curvature at fixed search rank **69**
J.1 A Hessian certificate for the convex duality gap . . . 69
J.2 Affine feasibility and an explicit dual witness . . . 70
J.3 Many negative-curvature directions . . . 71
J.4 Isotropic perturbations and local gradient descent . . . 72

K Proofs for overparameterized Gram objectives **74**
K.1 A representation-specific landscape bound . . . 74
K.2 Gaussian tangent moments and the landscape bound . . . 75
K.3 Square search rank and the limits of normalized descent . . . 76

L Applications and exact feasibility **77**
L.1 MAX-CUT . . . 82
L.2 Boolean CSP moments . . . 82
L.3 Direct rounding with inconsistent moments . . . 83
L.4 The complete Boolean Lasserre moment constraints . . . 84
L.5 Penalty limits . . . 85
L.6 Uniform affine feasibility and refutation . . . 85
L.7 Lovász theta and a finite certified interval . . . 87
L.8 Continuous optimization examples . . . 90
L.9 A penalized semidefinite relaxation of K-means . . . 90

Neural parameterizations **92**

M Neural parameterizations and their hypotheses **92**
M.1 The normalized-head transfer . . . 92
M.2 A sufficient finite-horizon condition . . . 92
M.3 Unnormalized convex losses . . . 93
M.4 A classical certificate for regularized two-layer networks . . . 93

CIRCUIT EXECUTION AND PREDICTION COMPLEXITY

A PROOFS FOR THE SIMULATION CONSTRUCTION

This appendix supplies the proofs behind Theorem 2.1 and its two companions, Propositions A.5 and A.6, and closes with the resource accounting for general programs (Appendix A.8). Throughout, the simulation is exact: every arithmetic operation uses the formulas written below, and the encoded initial state is part of the construction rather than a free parameter.

A.1 THE SCALAR LIFT

The construction of Section 2 rests on one identity, which turns a unit gradient step at a Boolean state into an arbitrary polynomial update. The lemma below states it in the generality needed both for the circuit objective and for the bounded clock of Appendix A.8; in it, the polynomial q reads a bit without passing a derivative through it, and r carries the derivative to the coordinate being written.

Lemma A.1 (Boolean update lift)**.** *Let $A\colon \{0,1\}^N \to \{0,1\}^N$ have polynomial update rules $A_i(b) = R_i(b)$. Define $q(u) = 3u^2 - 2u^3$ and $r(u) = u - q(u)$, with q applied coordinatewise. For any polynomial V, let*

$$\Phi(z) = V(q(z)) + \sum_{i=1}^{N} r(z_i)\big[q(z_i) - R_i(q(z))\big].$$

At every Boolean vertex b,

$$\Phi(b) = V(b), \qquad \nabla\Phi(b) = b - A(b).$$

Unit-step PGD on $[0,1]^N$ therefore maps b to $A(b)$.

Proof. We evaluate the value term and the N correction terms separately at a Boolean vertex b, using the four identities $q(b_i) = b_i$, $q'(b_i) = 0$, $r(b_i) = 0$, and $r'(b_i) = 1$, which hold in every coordinate. Because $r(b_i) = 0$, every correction term vanishes at b, so $\Phi(b) = V(q(b)) = V(b)$. As for the gradient, every first-order partial derivative of $V(q(z))$ carries a factor $q'(b_i)$ and hence vanishes. In the ith correction term, derivatives of the bracket are multiplied by $r(b_i) = 0$, whereas differentiating the leading factor $r(z_i)$ leaves $r'(b_i)[q(b_i) - R_i(q(b))] = b_i - R_i(b)$ in coordinate i and nothing elsewhere. Summing over i gives $\nabla\Phi(b) = b - A(b)$, and the computation is unaffected when a rule reads its own target or uses a parent more than once, since only the leading factor of each term contributes. Finally, $b - \nabla\Phi(b) = A(b)$ already lies in the box $[0,1]^N$, so the projection leaves it unchanged and the unit PGD step maps b to $A(b)$. □

An input coordinate carries the identity rule $R_i(z) = z_i$, whose correction term $r(z_i)[q(z_i) - q(z_i)]$ is identically zero, and its partial derivatives through every other factor vanish at Boolean states because they carry a factor q'. Input coordinates may therefore either remain free coordinates of the unit cube or be clamped to their specified values without changing the trajectory.

A.2 RAILS AND FINITE-DELAY CIRCUITS

Theorem 2.1 applies the lift to the dual-rail encoding of Section 2, whose update rules we now write out in full. Every wire w has two coordinates, the false rail f_w and the true rail t_w, and the three valid states are unknown $(0,0)$, false $(1,0)$, and true $(0,1)$. Input wires are initialized to their specified Boolean values and internal wires to unknown, while a constant-one wire, if the circuit has one, is initialized to $(0,1)$ and carries the identity update like an input. Writing $u \vee v = u + v - uv$ for the polynomial OR, the rules for a gate g with parent wires a, b are

Gate	f_g^+	t_g^+
AND	$f_a \vee f_b$	$t_a t_b$
XOR	$(f_a f_b) \vee (t_a t_b)$	$(f_a t_b) \vee (t_a f_b)$

Each rule has degree at most four and reads at most four parent coordinates, and repeated parent wires are permitted.

Definition A.2 (Finite delay)**.** A circuit has delay at most D on a specified input if these synchronous rules resolve every internal wire within D rounds. Here an AND gate may resolve to false after learning only one false input, whereas an XOR gate requires both input values, and the directed circuit may contain cycles.

The decrease identity in Theorem 2.1 requires each gate to spend exactly one rail change, which Lemma A.3 supplies for the synchronous Boolean rail dynamics; the proof of Theorem 2.1 then identifies these dynamics with PGD through Lemma A.1.

Lemma A.3 (Monotone resolution)**.** *Along the trajectory of the synchronous rail rules from the initialization above, every rail is nondecreasing and each wire remains in one of the three valid states. In a complete finite-delay run, exactly one rail per gate changes from zero to one, and no rail changes more than once.*

Proof. Two inductions on the round index are needed, one for monotonicity and one for validity. Since the rail rules use only AND and OR applied to nonnegative Boolean rails, each rule is monotone in the coordinatewise order. The first updated internal state is at least the all-zero initial internal state and the inputs remain fixed, so the induction hypothesis $z^t \geq z^{t-1}$ together with monotonicity of the rules gives $z^{t+1} \geq z^t$ at every round.

Validity of each wire's state follows by a second induction. An AND output can acquire a true rail only when both parents are true, in which case neither parent has the false rail needed to assert the output's false rail. For XOR, the false and true rules fire on equal and on unequal known parent values respectively, which cannot both hold when each parent is in a valid state. A wire therefore never asserts both rails. Because the rules implement the stated propagation semantics and finite delay resolves every gate, each gate changes exactly one rail from zero to one and, by monotonicity, no rail changes more than once. □

A.3 Exact simulation and decrease

With the lift and the monotonicity lemma in place, the main theorem follows by assembling the objective and reading off its decrease along the trajectory.

Proof of Theorem 2.1. The proof has three parts: the trajectory, obtained from Lemmas A.1 and A.3; the degree and size count; and the decrease identity together with the work and word bounds. Applying Lemma A.1 with identity rules on the input rails and $V(z) = -\sum_{i \in I} z_i$ on the internal rails produces the objective (2), so at every reached Boolean state the unit PGD update is the synchronous rail update. Induction on the round then gives the claimed trajectory and its completion within D rounds. Once all gates have resolved, the state is stationary: monotonicity (Lemma A.3) gives $A(b) \geq b$ coordinatewise and validity forbids a second rail on a resolved gate, so $A(b) = b$ and $\nabla \Phi_C(b) = b - A(b) = 0$.

The degree and size count is direct. Since q and r have degree three and $\deg R_i \leq 4$, a correction term has degree at most $3 + 3 \cdot 4 = 15$ and involves its target together with at most four parents. Combining each internal rail's unary value term with its correction term yields one factor per internal rail. With $2S$ internal rails, $2B$ input rails, and at most two constant coordinates, the representation and its incidence lists have size $O(B + S)$.

To obtain the decrease identity, let E_t be the set of rails that change in round t; by Lemma A.3 each change is from zero to one. At a Boolean state the correction terms vanish and the objective equals $-\sum_{i \in I} z_i$, which by validity (one rail per resolved gate) is minus the number of resolved gates, so the objective changes by $-|E_t|$. The update identity gives a gradient of -1 on each changing rail and zero elsewhere. Hence

$$\Phi_C(z^{t+1}) - \Phi_C(z^t) = -|E_t| = -\|z^{t+1} - z^t\|_2^2 = -\|\nabla \Phi_C(z^t)\|_2^2.$$

The cache argument below (Lemma A.4) proves the total work and storage bounds. Since every reached argument is Boolean and all local formulas have bounded degree and fixed integer coefficients, local intermediate arithmetic is bounded. At Boolean states the accumulated gradient entries lie in $\{-1, 0, 1\}$, so the $O(\log(B + S))$-bit words are needed only for array indices, which gives the stated word bound. □

In plain terms, the theorem says that each round of the circuit is one gradient step whose squared length is the number of gates it resolves. An Armijo line search that tries the unit step first accepts it along this trajectory for every sufficient-decrease parameter in $(0, 1)$, although this is an assertion about the initialized trajectory and supplies no step-size bound at fractional states.

A.4 Counting the cost of a full simultaneous gradient

The work bound in Theorem 2.1 rests on a general accounting lemma for PGD on a sum of local factors, which charges every operation of an incremental evaluator to a coordinate change. It is stated here in a form that Appendix A.8 and Proposition A.6 reuse.

Lemma A.4 (Cached factor evaluation)**.** *Let $\Phi(z) = \sum_{a=1}^{M} \phi_a(z_{I_a})$ be a sum of M factors on a product box in $\mathbb{R}^N$, where I_a is the set of coordinates read by factor a, and suppose that evaluating a factor and all its partial derivatives costs $c_a \geq \max\{1, |I_a|\}$ operations. If coordinate i changes κ_i times before the PGD trajectory becomes stationary, exact simultaneous PGD and its objective values can be evaluated in total work*

$$O\left(N + \sum_i \kappa_i + \sum_a c_a \left[1 + \sum_{i \in I_a} \kappa_i\right]\right).$$

When factor arity and cost are constant and each coordinate changes a constant number of times, the total work is $O(N + M)$.

Proof. The evaluator caches every factor and touches only what changes; we describe its data structures, verify that the cache stays correct, and then charge each operation to a coordinate change. Store the current vector, every factor value with its partial derivatives, the gradient sums $g_i = \sum_{a:\, i \in I_a} \partial_i \phi_a$, and reverse incidence lists connecting each coordinate to the factors that read it. Maintain in addition the active set of coordinates whose projected displacement $d_i = \Pi_{K_i}(z_i - g_i) - z_i$ is nonzero. Initialization evaluates all factors and builds these structures.

In each round, save every active displacement and then commit all saved changes simultaneously, so that every evaluation reads the old state. Reevaluate each factor incident to a changed coordinate, marking factors so that none is evaluated twice in a round, and update the gradient sums by subtracting the old derivatives and adding their replacements. Then refresh the active-set membership of every coordinate that changed or received a changed gradient contribution; the first condition is needed because the displacement depends on the coordinate's own value as well as on its gradient.

Correctness follows because an unchanged factor keeps the same arguments, so its cached derivatives remain valid, and induction on rounds gives agreement with full simultaneous PGD. As for the cost, factor a is refreshed at most $\sum_{i \in I_a} \kappa_i$ times, and each refresh's evaluation cost $c_a \geq |I_a|$ pays for the associated incidence traversal and derivative updates. Saving and committing changed coordinates costs $O(\sum_i \kappa_i)$ in total. Since every active round changes at least one coordinate, there are at most $\sum_i \kappa_i + 1$ rounds, the last of which finds the active set empty, so the remaining per-round bookkeeping is charged to these changes as well. Membership flags and array positions support constant-time insertion into and deletion from the active set, and maintaining the sum of factor values costs no more than maintaining the derivatives, which completes the accounting. □

For the circuit construction, $N = O(B + S)$, $M = O(S)$, every rail changes at most once, and each factor has constant arity and cost, so the lemma gives the $O(B + S)$ work bound of Theorem 2.1. The argument tolerates unbounded fanout because the total number of factor incidences remains $O(S)$. The evaluator stops when the active set is empty, and a caller who wants dense copies of every iterate pays their output cost in addition.

A.5 A strict local minimum at the terminal state

The terminal state of Theorem 2.1 is stationary, and Section 2 promised that a small change to the value term makes it a strict local minimum; this subsection supplies that change.

Proposition A.5. *Clamp the input rails and replace the value term in (2) by $\widetilde{V}(q(z))$, where*

$$\widetilde{V}(f, t) = \sum_{w \text{ internal}} (-f_w - t_w + 2 f_w t_w).$$

The trajectory, work bound, and decrease identity are unchanged, and at the terminal state the Hessian on the internal coordinates is $6I$, so the terminal state is a strict local minimum.

Proof. We check that the new value term leaves the trajectory unchanged and then compute the Hessian at the terminal state term by term. On a valid wire state $f_w t_w = 0$, so $\widetilde{V}$ is still minus the number of resolved gates, and its composition with q has zero gradient at every Boolean state because each first partial carries a factor q'. Lemma A.1 therefore gives the same update, hence the same trajectory and the same decrease identity, while the work bound survives because the modification adds only constant-arity factors per gate.

At the terminal state every correction residual $q(z_i) - R_i(q(z))$ is zero, its first derivatives vanish because they contain q', and its leading factor $r(z_i)$ vanishes as well. Two applications of the product rule then show that every correction term has zero Hessian there. In $\widetilde{V}(q(z))$, the mixed second derivatives contain two factors of q' and vanish, whereas for a resolved pair $(f, t) = (1, 0)$ the two diagonal entries are $(-1+2q(t))q''(f) = 6$ and $(-1+2q(f))q''(t) = 6$, and the pair $(0, 1)$ gives the same values. The Hessian on the free coordinates is therefore $6I$ at a point where the free-coordinate gradient is zero, and the second-derivative test yields a strict local minimum. □

In effect, the modified objective turns the computed answer from a stationary point into a basin, although only for the terminal state reached from the encoded initialization: nothing is claimed about convergence from arbitrary initializations or about stability under a perturbation of specified size.

A.6 Worked example with a feedback edge

A two-gate circuit with a feedback edge illustrates that finite delay, and not acyclicity, is what the construction needs. Consider the two gates

$$g = x \text{ AND } h, \qquad h = g \text{ XOR } 1,$$

with input $x = 0$. Although the graph has the cycle $g \to h \to g$, the known false input resolves g in the first round, after which the second round resolves h. In the coordinate order (f_g, t_g, f_h, t_h),

Round	Internal rails	Resolved gate values	Φ_C
0	$(0, 0, 0, 0)$	none	0
1	$(1, 0, 0, 0)$	$g = 0$	-1
2	$(1, 0, 0, 1)$	$g = 0,\ h = 1$	-2

After clamping x and the constant wire, the update rules simplify to $R_{f_g} = 1$, $R_{t_g} = 0$, $R_{f_h} = t_g$, and $R_{t_h} = f_g$, and writing $z = (z_1, z_2, z_3, z_4)$ gives the explicit objective

$$\begin{aligned}\Phi_C(z) = &-\sum_{i=1}^{4} q(z_i) + r(z_1)[q(z_1) - 1] + r(z_2)q(z_2) \\ &+ r(z_3)[q(z_3) - q(z_2)] + r(z_4)[q(z_4) - q(z_1)].\end{aligned}$$

Its first two gradients are $(-1, 0, 0, 0)$ and $(0, 0, 0, -1)$ and its terminal gradient is zero, so the two active steps each reduce the objective by one, as Theorem 2.1 asserts.

A.7 Message passing

Because the objective is a sum of local factors, its gradient step can be computed by message passing on the variable–factor graph, which is the implementation promised in Section 2.

Proposition A.6 (Exact realization on reached states)**.** *The PGD circuit construction has a fixed-weight ReLU message-passing realization with $O(B + S)$ nodes and edges and constant state and message dimension, in which two communication phases implement each PGD step. Total work is $O(B + S)$ with incremental evaluation and $O((B + S)(D + 1))$ with full graph evaluation in every round.*

Proof. We build the graph, show that two phases compute one PGD step, and then realize each local derivative table by a small ReLU network. Make one node for each coordinate and one for each factor, with an edge for every incidence; a finite node type identifies the polynomial template and finite edge labels identify argument positions. In the first phase, variable nodes send their current values to adjacent factor nodes, and in the second phase each factor returns its partial derivative to each incident variable. Summing the returned values gives the full gradient, after which each variable applies its coordinate projection, so the states at phase boundaries are the PGD iterates.

It remains to realize each local derivative table by a ReLU network. For $d \le 5$ arguments and $a \in \{0,1\}^d$, let the bump

$$\beta_a(x) = \mathrm{ReLU}\left(1 - \sum_{j=1}^{d} |x_j - a_j|\right).$$

On the Boolean cube β_a equals one at a and zero at every other vertex, so any local table p is represented on the reached states by $\sum_{a\in\{0,1\}^d} p(a)\beta_a(x)$, with a constant number of terms and a constant number of derivative outputs. Projection is the ReLU formula $\Pi_{[0,1]}(s) = \mathrm{ReLU}(s) - \mathrm{ReLU}(s-1)$.

To obtain the work bound, store messages and aggregate sums separately for the two phases, and when a variable changes notify only its incident factors and replace their stored derivative messages. All evaluations are completed against the old state before the next variable state is committed, so Lemma A.4 gives $O(B+S)$ total work, whereas evaluating the full graph in every phase multiplies this cost by $O(D+1)$. □

Put differently, the compiled variable–factor graph runs the descent by itself with constant-size messages. The ReLU formulas match the polynomial derivatives only on Boolean arguments, which suffices for this exact-trajectory result, and the graph is the compiled one. Realizing the same computation on a prescribed input graph would be a separate communication problem.

A.8 Computational models and resource bounds

The simulation results of this paper represent a computation in different ways, and this subsection records the resource bounds of each so that they are not compared out of context. The finite-delay construction expands the source program into wires that resolve once, whereas the recurrent clock constructions reuse coordinates that change many times, and each guarantee carries its own space, precision, and initialization conditions.

General programs through circuits. For a bounded-word PRAM in the model of Heath (2023), let n be its input-size parameter, B its encoded input bit count, $W(n)$ a supplied polynomial worst-case work bound, and T the parallel time of the specified execution. Its words have $\Theta(\log n)$ bits, and its primitive operations satisfy the compiler's circuit-size and delay requirements, so the compiler produces $S = O(W(n)\log^4 n)$ gates with delay $D = O(T\log^3 n)$. If a gate description can be generated in time $u(n) = \mathrm{polylog}(n)$, then generating the objective and executing PGD takes $O(B + W(n)\log^4 n\, u(n))$ work, while the representation and the cached execution use $O(B+W(n)\log^4 n)$ space and work respectively. Because the supplied budget may exceed the work of an easy input, the bound keeps its dependence on $W(n)$.

A bounded clock for compact local programs. For completeness, the same lift gives a particularly simple reusable Boolean construction. Suppose A updates d Boolean state bits by constant-arity rules R_i and has an absorbing halting state, in the sense that $A(b) = b$ whenever the halt bit satisfies $b_h = 1$, and suppose the specified execution halts within L transitions. With the correction sum Λ and a clock coordinate c, let

$$\Lambda(z) = \sum_{i=1}^{d} r(z_i)[q(z_i) - R_i(q(z))], \qquad \Phi_{\mathrm{loc}}(z,c) = \Lambda(z) + dc(1 - q(z_h)).$$

Use the fixed box $[0,1]^d \times [0, dL]$ and initialize $c = dL$. At a reached Boolean state b before halting,

$$\nabla_z \Phi_{\mathrm{loc}} = b - A(b), \qquad \partial_c \Phi_{\mathrm{loc}} = d,$$

so each unit PGD step performs one source transition and decreases c by d. Once $b_h = 1$, the clock term has zero gradient in both z and c because $q'(1) = 0$ and $1 - q(1) = 0$, and the correction gradient $b - A(b)$ vanishes by the absorbing hypothesis, so the entire state is stationary. Each active step decreases the objective by at least d^2, because at a Boolean state the objective equals $dc(1 - b_h)$ and c falls by d, while its squared gradient norm is at most $d^2 + d \le 2d^2$. There are $O(d)$ constant-arity factors, so if bit i changes κ_i times and occurs in δ_i factors, Lemma A.4 bounds the work by $O(d + L + \sum_i \kappa_i \max\{1, \delta_i\})$, and for an execution of length $\ell \le L$ the additive L can be replaced by ℓ. The clock uses $O(\log(dL))$ bits.

The strongly convex recurrent clock. Theorem C.1 gives a different reusable-state guarantee, in which the local clocks shrink geometrically while the objective remains globally strongly convex with fixed conditioning. The horizon is absent from the energy's description and appears only in the readout margin; Appendix C supplies the proof and the arithmetic representation.

Randomness and continuation. Fixing a finite random tape reduces a bounded randomized execution to a deterministic program, for which the exact-output theorem holds pathwise. For an unbounded random execution, a finite-memory simulator must also specify how it receives further random bits. The word-level construction uses a succinct C^1 piecewise-polynomial description, exact integer state coordinates, and externally connected PGD segments: between segments it retains the unfinished program state and the unused random bits, refills the buffer, and resets its clock. Under the declared constant-word instruction model, its peak representation has $O(S_{\mathrm{mem}} + P + 1)$ words, where P counts the program description and S_{mem} the live memory, and a terminating execution of T_{src} instructions costs $O(S_{\mathrm{mem}} + P + T_{\mathrm{src}})$ word operations. Because the external resets can increase the objective, these bounds describe a continued simulator composed of PGD segments rather than one autonomous trajectory or one explicitly listed global polynomial.

B Circuit evaluation at a strongly convex minimizer

Theorem 2.1 stored a computation in the trajectory of a nonconvex objective. This appendix proves Proposition 2.2, which stores a circuit's answer at the unique minimizer of a strongly convex one, with its numerical and initialization conditions stated explicitly. Two devices carry the construction: a flat gate interpolation removes backward derivatives at Boolean states, and coordinate scales that decrease along dependency edges control the global Hessian, including at points outside those states.

B.1 Circuit normalization and coordinate scales

Let V_0 be the set of clamped input sites and V_1 the set of gates whose states are optimized, and write $n = |V_1|$ and $S_{\mathrm{tot}} = |V_0| + n$ for the total circuit size, which therefore includes the input sites. The circuit is a directed acyclic graph over AND, OR, NOT, and COPY gates with at most two parents per gate, and constant inputs may be included among the clamped sites. Two normalizations are applied: sites with no path to a designated output are pruned, and an AND or OR gate with two identical parents is replaced by a COPY gate, so that each parent set P_v consists of distinct sites. Write $v \to w$ when v is a parent of w.

The input $b \in \{0,1\}^{V_0}$ is held fixed. Input sites have depth zero, the depth of a gate is one plus the largest depth of its parents, and $D \ge 1$ denotes the maximum gate depth. The count $r \ge 1$ is the number of sinks; after pruning every sink is an output, and an output that is itself an input is read directly from b. An arbitrary two-input Boolean truth table can be expanded into the stated basis with constant size and depth overhead. Throughout this appendix, S_{tot}, D, and r refer to the circuit after pruning and after the replacement of AND/OR gates with identical parents by COPY gates.

Let $z^*(b) \in \{0,1\}^{V_1}$ be the correct gate-value vector for the input b. Positive dyadic scales a_v, which depend only on the circuit and not on b, will place this vector at $x^*_v = a_v z^*_v(b)$, and an output bit at gate o is decoded by the test $x_o > a_o/2$.

The Hessian bound below needs the squared scale ratios along the edges leaving each site to sum to a small budget θ, while the scales must not become too small for the numerical description. The lemma constructs such scales for arbitrary fanout.

Lemma B.1 (Dyadic scales for arbitrary fanout). *For every $0 < \theta \le 1$, there are integer powers of two $a_v \in (0, 1]$ such that*

$$\sum_{w:v\to w} (a_w/a_v)^2 \le \theta \quad \textit{for every site } v. \tag{11}$$

Their largest scale exponent $P = \log_2(1/\min_v a_v)$ satisfies

$$P \le \tfrac{1}{2}\log_2 r + \tfrac{D}{2}\log_2(8/\theta). \tag{12}$$

Proof. The construction assigns integer weights in reverse topological order and then normalizes them, and the bound on P comes from counting paths to sinks. Give each sink weight $W_v = 1$, and in reverse topological order give each nonsink the least power of four that is at least $\theta^{-1}\sum_{w:v\to w} W_w$. The weights are positive integer powers of four, so with $W_{\max} = \max_v W_v$ and $a_v = \sqrt{W_v/W_{\max}}$ the scales are powers of two in $(0, 1]$, and (11) holds because $\sum_{w:v\to w}(a_w/a_v)^2 = W_v^{-1}\sum_{w:v\to w} W_w \le \theta$.

The exponent bound uses that $\theta \le 1$ and every weight is at least one, so the quantity $\theta^{-1}\sum_{w:v\to w} W_w$ is at least one at every nonsink. Rounding it upward to a power of four therefore at most quadruples it, and

$$W_v \le (4/\theta)\sum_{w:v\to w} W_w$$

at every nonsink. Repeated substitution bounds W_v by the sum of $(4/\theta)^{|\pi|}$ over all directed paths π from v to sinks, each of length at most D. To count these paths for a fixed sink, unfold its parent dependencies into a binary tree and stop each branch when it reaches v. The tree has depth at most D, hence at most 2^D leaves labeled v, and these leaves are in bijection with the paths from v to that sink. Summing over the r sinks gives

$$W_v \le r\,2^D(4/\theta)^D = r(8/\theta)^D.$$

Every sink has weight one, and a sink exists because the graph is finite and acyclic, so $\min_v a_v = W_{\max}^{-1/2}$, and taking logarithms proves (12). □

For the remaining construction fix $\theta = 1/4096$, so that $\log_2(8/\theta) = 15$ and (12) reads $P \le 7.5D + \frac{1}{2}\log_2 r$, regardless of the circuit's maximum fanout.

B.2 The local energy

The energy is assembled from two scalar functions, which we fix first, and one gate polynomial per optimized site. Define the clipped scalar interpolation p, which reads a parent value, and the bounded output extension H, which reads a gate's own output, by

$$p(t) = \begin{cases} 0, & t \le 0, \\ 3t^2 - 2t^3, & 0 \le t \le 1, \\ 1, & t \ge 1, \end{cases} \qquad H(t) = \begin{cases} -4, & t \le -8, \\ t + t^2/16, & -8 \le t \le 0, \\ t, & 0 \le t \le 1, \\ t - (t-1)^2/16, & 1 \le t \le 9, \\ 5, & t \ge 9. \end{cases} \tag{13}$$

Their values and first derivatives agree at every breakpoint, so both functions are globally $C^{1,1}$ with

$$|p'| \le 3/2, \quad |p''| \le 6, \quad |H| \le 5, \quad |H'| \le 1, \quad |H''| \le 1/8, \tag{14}$$

where second derivatives here and throughout this appendix are interpreted almost everywhere. At the Boolean points $p(z) = z$, $p'(z) = 0$, $H(z) = z$, and $H'(z) = 1$, which is what makes a unit gradient step act as an exact gate evaluation.

Write $y_v = x_v/a_v$ for the normalized value of each optimized coordinate. For gate v, let $G_v(y; b)$ be its gate polynomial with every optimized parent value y_i replaced by $p(y_i)$ and every clamped parent replaced by b_i. Here AND uses the product of its two parent values, OR their sum minus their product, and NOT and COPY the complement and the identity; since p takes values in $[0, 1]$, so does G_v. Differentiating these gate polynomials and applying (14) bounds each first partial

by $3/2$, each pure second partial by 6, and each mixed second partial by $(3/2)^2 = 9/4$. A first partial vanishes whenever the differentiated parent coordinate is Boolean because p' does. The distinct-parent convention matters here: with a repeated parent, an AND gate would read $p(y_i)^2$, whose derivative $2p(y_i)p'(y_i)$ exceeds $3/2$ on part of $(0,1)$ and is bounded by the scalar bounds only through $2 \cdot (3/2) = 3$.

Given the clamped input b, define the energy

$$E_b(x) = \frac{1}{2}\sum_{v\in V_1} x_v^2 - \sum_{v\in V_1} a_v^2 H(x_v/a_v) G_v(y;b), \tag{15}$$

whose first sum is quadratic in each coordinate and whose second sum contains one interaction factor per optimized gate. Each factor involves one output and at most two optimized parents, the input enters only through its clamped slots, and there are $O(S_{\mathrm{tot}})$ factors, each with a constant number of polynomial pieces of degree at most eight.

The main statement of this appendix collects the global regularity, the exact trajectory, and the numerical description of this energy.

Theorem B.2 (The ordinary-gradient circuit compiler)**.** *The energy (15) is globally $C^{1,1}$, $3/4$-strongly convex, and $5/4$-smooth. Its unit-GD map is globally $1/4$-contractive, and*

$$E_b(x - \nabla E_b(x)) \le E_b(x) - \tfrac{3}{8}\|\nabla E_b(x)\|_2^2. \tag{16}$$

From every initialization $x^0 = a \odot z^0$ with $z^0 \in \{0,1\}^{V_1}$, it reaches the unique global minimizer $x^* = a \odot z^*(b)$ within D iterations. At this point,

$$\nabla E_b(x^*) = 0, \qquad E_b(x^*) = -\tfrac{1}{2}\sum_{v:z_v^*(b)=1} a_v^2. \tag{17}$$

Each local numerical coefficient, breakpoint, and exact orbit coordinate has an $O(D + \log r)$-bit rational description. Dense evaluation of the D steps uses $O(S_{\mathrm{tot}} D)$ arithmetic operations.

In words, the energy is a well-conditioned strongly convex function whatever the input, and gradient descent from any scaled Boolean start evaluates the circuit level by level, reaching the minimizer within depth-many steps.

B.3 Global conditioning

The argument begins with the Hessian bound, which is the only place where the scale budget θ enters, and then integrates it.

Proof of Theorem B.2. We proceed in four steps: global regularity; the Hessian bound (18); the integrated inequalities that yield strong convexity, smoothness, contraction, and decrease; and the exact trajectory with its numerical and work accounting. For regularity, the scalar functions H and p and their first derivatives are bounded, and their first derivatives are globally Lipschitz. For fixed positive scales, finite products of their scaled versions therefore have globally Lipschitz first derivatives, so each interaction factor, and with it E_b, is globally $C^{1,1}$. The pieces are separated by finitely many coordinate hyperplanes, and the largest factor degree is $2 + 3 + 3 = 8$.

At a point off these hyperplanes, $\nabla^2 E_b - I$ decomposes into exactly three parts according to which arguments of a factor are differentiated: the output–output diagonal, the output–parent block $-(M + M^\top)$ defined below, and the parent–parent block. We bound each part and then sum the bounds. Differentiating a factor twice in its own output gives a diagonal matrix with entries $-H''(y_v)G_v$, whose operator norm is at most $1/8$ by (14) and $0 \le G_v \le 1$.

For the output–parent derivatives, define a matrix M by

$$M_{vi} = (a_v/a_i)H'(y_v)\partial_i G_v \quad (i \in P_v \cap V_1),$$

with all other entries zero, where $\partial_i G_v$ differentiates the normalized parent variable. Fan-in at most two, the bounds $|H'| \le 1$ and $|\partial_i G_v| \le 3/2$, and the scale budget (11) give

$$\begin{aligned}\|Mu\|_2^2 &\le 2(3/2)^2 \sum_v \sum_{i\in P_v\cap V_1} (a_v/a_i)^2 u_i^2 \\ &\le (9/2)\theta\|u\|_2^2,\end{aligned}$$

so the output–parent correction $-(M + M^\top)$ has norm at most $2\|M\|_{\mathrm{op}} \le 3\sqrt{2\theta}$.

The parent–parent block differentiates the gate function twice. Using $|H| \le 5$ and the pure second-partial bound 6, the pure second derivatives contribute at most

$$5 \cdot 6 \sum_v \sum_{i \in P_v \cap V_1} (a_v/a_i)^2 u_i^2 \le 30\theta \|u\|_2^2$$

to the absolute value of its quadratic form. If gate v has two optimized parents i, j, the mixed bound $9/4$ together with $2|st| \le s^2 + t^2$ shows that its mixed term is at most

$$2 \cdot 5 \cdot \frac{9}{4} \left| \frac{a_v}{a_i} u_i \frac{a_v}{a_j} u_j \right| \le \frac{45}{4} \big((a_v/a_i)^2 u_i^2 + (a_v/a_j)^2 u_j^2 \big).$$

Summing over gates and using (11) bounds this part by $(45/4)\theta\|u\|_2^2$. Since all three matrices are symmetric, their quadratic-form bounds also bound their operator norms, and adding the three bounds gives

$$\|\nabla^2 E_b - I\|_{\mathrm{op}} \le \frac{1}{8} + 3\sqrt{2\theta} + \frac{165}{4}\theta < \frac{1}{4}, \qquad \theta = 1/4096, \tag{18}$$

where for the chosen θ the right side is smaller than $1/8 + 3/32 + 1/64 = 15/64$.

The pointwise bound is now integrated. For a line segment contained in none of the break hyperplanes, integrate the Hessian along its finitely many smooth pieces, which continuity of the gradient joins. Approximating an arbitrary segment by such segments extends the estimates to every segment. The spectral bounds in (18), which place the Hessian between $\frac{3}{4}I$ and $\frac{5}{4}I$, give

$$E_b(y) \ge E_b(x) + \langle \nabla E_b(x), y - x \rangle + \tfrac{3}{8}\|y - x\|_2^2$$

and $\|\nabla E_b(y) - \nabla E_b(x)\|_2 \le (5/4)\|y - x\|_2$, while integrating $I - \nabla^2 E_b$ along the segment gives

$$\|(y - \nabla E_b(y)) - (x - \nabla E_b(x))\|_2 \le \tfrac{1}{4}\|y - x\|_2.$$

These are the stated strong convexity, smoothness, and contraction, and applying the smoothness inequality to $y = x - \nabla E_b(x)$ gives

$$E_b(y) \le E_b(x) - \|\nabla E_b(x)\|_2^2 + \tfrac{5}{8}\|\nabla E_b(x)\|_2^2,$$

which is (16).

Exact execution and the minimizer. Differentiating every factor in which x_v appears, including those of the downstream gates w with $v \to w$, gives

$$\partial_v E_b(x) = x_v - a_v H'(y_v) G_v - \sum_{w : v \to w} \frac{a_w^2}{a_v} H(y_w) \partial_v G_w. \tag{19}$$

When all normalized optimized coordinates are Boolean, $H'(y_v) = 1$ and every derivative $\partial_v G_w$ in the downstream sum vanishes because p' does, so unit GD satisfies

$$x_v - \partial_v E_b(x) = a_v \, \mathrm{gate}_v(y_{P_v}; b). \tag{20}$$

In words, each step evaluates every gate once on its parents' current values, and the scaled Boolean set is invariant. After one update every depth-one gate is correct because its parents are clamped inputs. If every gate of depth at most t is correct, then the next update makes every gate of depth at most $t + 1$ correct while leaving the already correct gates correct, since their parents remain correct. Induction on depth shows that all gates are correct by time D from every scaled Boolean starting state.

At that state, (20) is a fixed-point identity, so $\nabla E_b(x^*) = 0$, and strong convexity makes x^* the unique global minimizer. Substituting $H(z_v^*) = z_v^*$ and $G_v(z^*; b) = z_v^*$ into (15) gives (17).

Numerical descriptions and work. Every scale is a power of two with exponent at most P. The coefficients of a local degree-eight piece are fixed dyadic constants multiplied by a bounded number of scales or their reciprocals, the breakpoints are fixed dyadic constants times scales, and the orbit coordinates are zero or a_v. By Lemma B.1, each such numerical datum has $O(P+1) = O(D+\log r)$ bits. This is a statement about individual local numerical data; graph indices have their usual $O(\log S_{\mathrm{tot}})$-bit cost.

Since a factor has at most three optimized arguments, constant degree, and a constant number of pieces, evaluating its gradient and adding its contributions takes a constant number of arithmetic operations and comparisons. With $O(S_{\mathrm{tot}})$ factors, one dense GD step costs $O(S_{\mathrm{tot}})$ operations and D steps cost $O(S_{\mathrm{tot}}D)$, which is the stated work bound and completes the proof. This bound counts arithmetic operations, so finite arithmetic additionally charges the operand bit lengths, and the construction offers no analogue of the linear total work bound of the finite-delay compiler in Appendix A. □

Three remarks on the theorem's scope close this appendix. The global contraction bound holds from every real starting point, whereas the exact D-step guarantee uses the scaled Boolean initialization. Because the output amplitude a_o can be exponentially small in the circuit depth, the numerical description and the decoding margin $a_o/2$ remain part of the guarantee. The objective is globally $C^{1,1}$ and piecewise polynomial, with no higher regularity asserted at the breakpoints.

C Recurrent computation in a fixed convex energy

Unrolling a recurrent computation stores a separate state for each time step. To remove that storage cost, this appendix gives each Boolean register a real amplitude that decreases during the computation: the register value is the direction of a pair of coordinates, while gradient descent reduces their common scale. One fixed energy can therefore simulate every finite prefix of a recurrent computation, even though its unique minimizer contains no answer. Energy-dissipating recurrent simulation has longstanding precedents, including Šíma & Orponen (2001); what the result proved here adds is the specification of ordinary discrete gradient descent, constant conditioning, local factors, and the arithmetic needed for this particular representation.

C.1 The recurrent model and the local-clock theorem

The construction starts from a supplied local transition rule, so its cost must be measured before interpreting it as an implementation of an MPNN. Let $m \geq 1$ denote the number of Boolean registers and $F : \{0,1\}^m \to \{0,1\}^m$ their synchronous transition map. For each register i, the parent list P_i is an ordered list of at most two parent registers and F_i is an arbitrary Boolean truth table on those input occurrences, with constants, self-inputs, repeated inputs, and directed cycles all permitted. Let $S_i := \{i\} \cup \{j : j \text{ occurs in } P_i\}$ be the scope of factor i, the set of registers it uses, and define the maximum factor incidence

$$\Delta := \max_{1 \leq j \leq m} |\{i : j \in S_i\}|,$$

so that $1 \leq \Delta \leq m$. This incidence counts a factor once for each distinct register it contains, although the formulas below still count repeated input occurrences when taking products and derivatives.

To represent a register at a changing physical scale, use a signal $x_i \in \mathbb{R}$ and a clock $c_i \in \mathbb{R}$, and write $u = (x, c) \in \mathbb{R}^{2m}$ for the full physical state. At a synchronized positive clock value c, the Boolean configuration z is represented by $x_i = cz_i$ and $c_i = c$ for every i; the clocks are coordinates of the autonomous descent algorithm rather than an externally supplied time schedule.

Theorem C.1 (Local clocks for recurrent Boolean computation)**.** *For the transition map F above, fix the universal constant $K_0 := 200000$ together with the scale exponent k and the scale ratio τ given by*

$$k := \left\lceil \log_2\big(4(1 + K_0\Delta)\big) \right\rceil, \qquad \tau := 2^{-k}.$$

There is an explicit energy $E_F : \mathbb{R}^{2m} \to \mathbb{R}$, independent of the initial Boolean configuration and the time horizon, with the following properties.

1. *The energy is a sum of $O(m)$ factors, each involving at most six real coordinates; it is globally $C^{1,1}$ and $3/4$-strongly convex, has $5/4$-Lipschitz gradient, and its unique minimizer is the origin.*

2. *Ordinary gradient descent with step size one, initialized at $x_i^0 = z_i^0$ and $c_i^0 = 1$ for any $z^0 \in \{0,1\}^m$, satisfies*

$$c_i^t = \tau^t, \qquad x_i^t = \tau^t (F^t(z^0))_i \quad (i = 1, \dots, m;\ t \geq 0). \tag{21}$$

   The Boolean register at any finite time T is decoded by $x_i^T - c_i^T/2 > 0$.

3. *The unit-step map $u \mapsto u - \nabla E_F(u)$ is globally $1/4$-Lipschitz, and every exact update $u^+ = u - \nabla E_F(u)$ obeys*

$$E_F(u^+) \leq E_F(u) - \frac{3}{8}\|u^+ - u\|_2^2. \tag{22}$$

4. *Each update uses a gather stage and a return stage on the gate incidence graph, followed by local aggregation, and takes $O(m)$ arithmetic operations on explicit, constant-size piecewise rational factor formulas, whose bit cost is additional.*

The dimension $2m$ does not grow with the horizon, and the price paid for this is a decoder margin $\tau^T/2$ that shrinks with the horizon rather than a condition number that grows with it. The theorem therefore concerns finite-time computation: the answer is decoded from the transient, while the minimizer itself is the origin and carries no information. The proof of Theorem C.1 has three parts. Each gate is interpolated so that its input derivatives vanish at Boolean states; the local factors are then controlled globally, including at zero clocks, which establishes convexity; and a direct gradient calculation on synchronized Boolean states finally proves the orbit.

C.2 Gate interpolation and the clock factors

The first part of the proof is the gate interpolation, which must preserve truth-table values while preventing derivatives from feeding back through Boolean inputs. Define the clipped smoothstep $p : \mathbb{R} \to [0,1]$ by

$$p(t) := \begin{cases} 0, & t \leq 0, \\ 3t^2 - 2t^3, & 0 \leq t \leq 1, \\ 1, & t \geq 1, \end{cases}$$

which is $C^{1,1}$ with $|p'| \leq 3/2$ and $|p''| \leq 6$ almost everywhere, and which has $p'(0) = p'(1) = 0$. For a two-input truth table with values $b_{00}, b_{01}, b_{10}, b_{11} \in \{0,1\}$, let g be its multilinear extension

$$g(a,b) := b_{00} + (b_{10} - b_{00})a + (b_{01} - b_{00})b + (b_{11} - b_{10} - b_{01} + b_{00})ab,$$

with unary and constant gates using the corresponding affine and constant extensions, and define the interpolated gate $G_i(z) := g_i(p(z_j) : j \text{ occurs in } P_i)$. Because the multilinear extension is a convex combination of truth-table values on the unit square, $0 \leq G_i \leq 1$ on all of $\mathbb{R}^m$, while on the Boolean cube $G_i = F_i$ and every input derivative vanishes. A repeated parent contributes both chain-rule terms, each of which still contains p'.

The clock derivative needs a second cancellation in addition to flat gate inputs. Define the bounded interpolant $H : \mathbb{R} \to \mathbb{R}$ by

$$H(t) := \begin{cases} -1/2, & t \leq -1, \\ t + t^2/2, & -1 \leq t \leq 0, \\ t - 3t^2/2 + t^3, & 0 \leq t \leq 1, \\ 1/2 + (t-1) - (t-1)^2/2, & 1 \leq t \leq 2, \\ 1, & t \geq 2, \end{cases}$$

whose pieces match in value and first derivative, so that H is $C^{1,1}$ with $|H| \leq 1$, $|H'| \leq 1$, and $|H''| \leq 3$ almost everywhere, and whose derivatives are supported in $[-1,2]$. For a Boolean argument $z \in \{0,1\}$ it satisfies

$$H(z) = z/2, \qquad H'(z) = 1, \qquad 2H(z) - zH'(z) = 0, \tag{23}$$

and the third identity is the cancellation that will remove the own-clock derivative on Boolean states.

Because incomparable local clocks would make derivatives of their ratios unbounded, interactions between them are suppressed by the clock bump $\chi : \mathbb{R} \to [0, 1]$, defined by

$$\chi(r) := \begin{cases} p(2r-1), & r \le 1, \\ p(2-r), & r \ge 1. \end{cases}$$

This $C^{1,1}$ function is supported in $[1/2, 2]$ and satisfies $\chi(1) = 1$, $\chi'(1) = 0$, $|\chi'| \le 3$, and $|\chi''| \le 24$ almost everywhere; its vanishing derivative at one will leave synchronized clocks unchanged by the interaction terms.

For nonzero involved clocks, write $z_j := x_j/c_j$ for the normalized signal and define the local interaction

$$V_i(x, c) := c_i^2 H(z_i) G_i(z) \prod_{j \text{ occurs in } P_i} \chi(c_i/c_j), \tag{24}$$

and set $V_i = 0$ if any clock indexed by S_i is zero. A self-input has a constant ratio factor $\chi(c_i/c_i) = 1$ when $c_i \neq 0$, so that factor can be omitted. The proposed energy is

$$E_F(x, c) := \frac{1}{2} \sum_{i=1}^{m} x_i^2 + \frac{1-\tau}{2} \sum_{i=1}^{m} c_i^2 - \tau \sum_{i=1}^{m} V_i(x, c). \tag{25}$$

These formulas are piecewise rational in physical coordinates; their values at zero clocks are part of the definition, and their differentiability there is the principal point of Lemma C.2.

C.3 Global regularity and the exact descent orbit

The gate interpolation fixes the Boolean updates. The second part of the proof controls the factors globally. A small coefficient in front of the interactions preserves the quadratic energy's conditioning only if the interaction Hessians are uniformly bounded. The lemma below supplies that bound uniformly over clock magnitudes, signs, and all physical signal values. Summing this bound over the factors meeting a coordinate gives the theorem's strong convexity and smoothness constants.

Lemma C.2 (Uniform regularity of one local interaction)**.** *Each factor V_i in (24), extended by zero as specified, is globally $C^{1,1}$ with a K_0-Lipschitz gradient, where $K_0 = 200000$, and its gradient is zero whenever an involved clock is zero. The bound holds for arbitrary truth tables, self-inputs, and repeated inputs.*

Proof. Product derivatives are bounded where clocks are nonzero, with the clock bumps controlling all denominators, and those estimates are then extended across zero clocks. Because χ, χ', and its almost-everywhere second derivative all vanish outside $[1/2, 2]$, each parent clock obeys

$$\frac{1}{2} \le \frac{c_i}{c_j} \le 2$$

whenever a nonzero value, first-derivative term, or second-derivative term occurs. Every active term therefore has clocks of the same sign and comparable magnitudes. Every unbounded normalized signal in a derivative is multiplied by a derivative of its own H or p, whose compact derivative supports bound that normalized signal by two in absolute value.

To obtain an explicit constant, expand the gate polynomial into monomials. The absolute sum of its coefficients is at most $1 + 1 + 1 + 2 = 5$, a bound attained by XNOR. A monomial of V_i, apart from its scalar coefficient, has the form $s^2 A$, where $s := c_i$ is the own clock and A is a product of at most five functions, each of absolute value at most one. These consist of one H, at most two p functions, and at most two clock bumps. For each physical coordinate pair a, b, every nonzero product derivative term satisfies the conservative individual-function bounds

$$|\partial_a f| \le \frac{12}{|s|}, \qquad |\partial_{ab} f| \le \frac{432}{s^2}. \tag{26}$$

These bounds concern factors within an active product term, since an isolated $p(x_j/c_j)$ need not satisfy them when its clock is incomparable with s.

The constants in (26) follow from the three one-variable functions. For $H(x/s)$, the first-derivative numerators are H' and $-zH'$, bounded by two, while the second-derivative numerators are H'', $-(H' + zH'')$, and $2zH' + z^2H''$, bounded by sixteen. For $p(x/c_j)$, its derivative support and clock comparability give bounds $3/|s|$ and $36/s^2$. For $\chi(s/c_j)$, the first bound is $12/|s|$. The largest second bound is $4(2 \cdot 2 \cdot 3 + 2^2 \cdot 24)/s^2 = 432/s^2$, and self-input ratio factors were omitted before differentiation.

Applying the product rule in the physical variables gives

$$|\partial_a A| \le \frac{60}{|s|}, \qquad |\partial_{ab} A| \le \frac{5 \cdot 432 + 20 \cdot 12^2}{s^2} = \frac{5040}{s^2}.$$

Since a Hessian entry of $s^2 A$ collects the terms $2A$, two copies of $2s\,\partial A$, and $s^2\partial^2 A$, it has absolute value at most $2+120+120+5040 = 5282$. A factor involves at most six physical coordinates and the coefficient absolute sum is at most five, so the row-sum (Gershgorin) bound $\|W\|_2 \le \max_i \sum_j |W_{ij}|$ for a symmetric matrix W gives

$$\|\nabla^2 V_i\|_2 \le 6 \cdot 5 \cdot 5282 = 158460 < K_0 \tag{27}$$

almost everywhere away from zero clocks. This estimate already includes repeated inputs, because their repeated p and χ factors occur among the at most five product factors, and identifying a self-input adds no multiplicity factor either, because the product rule was applied in physical coordinates throughout.

The same calculation controls the extension at zero clocks. Wherever the clocks are nonzero, $|V_i| \le s^2$ directly, and each gradient entry is bounded by $5(2 + 60)|s| = 310|s|$. If one involved clock is zero and another is nonzero, a clock bump vanishes on a neighborhood of that point. If all involved clocks are zero, the value is $O(s^2)$ and the gradient is $O(|s|)$ uniformly in the signals. Since $|s|$ is at most the distance to that zero-clock point, the zero extension is differentiable there with zero gradient, which is moreover continuous. Zero clocks with nonzero signals are covered by the same argument.

It remains to turn the almost-everywhere bound into a global one. On a generic line segment, each clock has at most one zero and each ratio or normalized-signal boundary has finitely many crossings, so the first derivatives are continuously piecewise rational between crossings. Integrating (27) on those pieces and using gradient continuity at the crossings gives the same Lipschitz bound on the whole segment. Exceptional segments follow by approximation with generic segments and gradient continuity, and a segment contained in an involved zero-clock hyperplane has zero factor gradient throughout. Hence V_i is globally $C^{1,1}$ with the claimed Lipschitz constant, and its gradient vanishes at zero clocks, as asserted. □

The factor estimate is independent of the number of registers. Summing it costs a factor equal to the number of factors meeting a coordinate, which is why the theorem is stated in terms of the distinct-factor incidence Δ.

Proof of Theorem C.1. Conditioning is established by summing the local Hessian bounds, after which the gradient is calculated on synchronized Boolean states. The choice of τ enters only in the first step. For a vector $v \in \mathbb{R}^{2m}$, write v_{S_i} for its signal and clock entries belonging to registers in S_i. Lemma C.2 gives

$$\left| v^{\mathsf{T}} \nabla^2 \Big(\sum_i V_i \Big) v \right| \le K_0 \sum_i \|v_{S_i}\|_2^2 \le K_0 \Delta \|v\|_2^2$$

almost everywhere, and comparing (25) with $\|u\|_2^2/2$ yields

$$\|\nabla^2 E_F - I\|_2 \le \tau(1 + K_0\Delta) \le \frac{1}{4}.$$

Because E_F is globally $C^{1,1}$, these almost-everywhere bounds integrate along segments, so E_F is $3/4$-strongly convex and $5/4$-smooth and $u \mapsto u - \nabla E_F(u)$ is $1/4$-Lipschitz. The smoothness inequality at step one gives (22).

For the exact orbit, suppose $c_i = c \neq 0$ and $x_i = cz_i$ with $z \in \{0,1\}^m$. All clock bumps then equal one with vanishing first derivatives, and all gate-input derivatives vanish as well, so the only surviving

signal derivative is

$$\partial_{x_i} \sum_j V_j = cH'(z_i)F_i(z) = cF_i(z).$$

Every clock derivative is zero: the input and ratio terms vanish, and the own-clock term is $c[2H(z_i) - z_iH'(z_i)]F_i(z) = 0$ by (23). This calculation includes self-inputs, because their gate-input derivatives are still zero. The full energy gradient is therefore

$$\partial_{x_i} E_F = c(z_i - \tau F_i(z)), \qquad \partial_{c_i} E_F = (1-\tau)c, \tag{28}$$

and subtracting it gives $(x_i^+, c_i^+) = (\tau c F_i(z), \tau c)$, so induction from $c = 1$ proves (21). The origin has zero gradient by the zero-clock extension, and strong convexity makes it the unique minimizer.

Finally, store (x_i, c_i) at register i. A factor gathers its parent pairs, evaluates its constant-size formula and derivatives, and returns each parent's derivative contribution; summing the incident contributions gives the gradient. Since at most three registers meet each factor, there are $O(m)$ such operations and incidences, which proves the locality and work statements, with arithmetic bit costs still to be charged. □

The proof supplies rational local modules, and implementing their arithmetic with finite-bit messages incurs an additional bit cost. Appendix D provides a rounded construction with an explicit precision guarantee.

D LOCAL ROUNDED GRADIENT COMPUTATION

The rational modules of Appendix C require a bit-cost analysis. The prediction theorem (Theorem 3.1) also needs a single local energy with explicitly specified finite arithmetic. The compiler below supplies both. Its proof establishes global regularity, the rounding invariant, and the local message modules. Within that proof, Corollary D.4 establishes exact Boolean execution by the same energy from the prescribed initialization.

Fix a synchronous Boolean transition map $F : \{0,1\}^m \to \{0,1\}^m$ in which each coordinate has at most two distinct parents, repeated parent occurrences having been collapsed into a unary truth table. Let S_i consist of register i and its parents, and assume that the maximum factor incidence satisfies $\Delta = \max_j |\{i : j \in S_i\}| \le 4$. The tape pipeline of Appendix E has $\Delta \le 3$, which discharges this hypothesis there. The dynamic coordinates are $X = (x, u) \in \mathbb{R}^{2m}$, and the fixed constants are the scale ratio $\tau = 2^{-25}$, the clock rescaling $\alpha = 64$, and the time-indexed scale $\gamma_t := \tau^t$.

The conventions differ from those of Appendix C in four places. The interpolant H below is a different function, affine on a $1/8$-neighborhood of each Boolean value, whereas the interpolant there satisfies the clock cancellation (23) only at the two Boolean points. The scale ratio is the fixed $\tau = 2^{-25}$ rather than $2^{-k(\Delta)}$. The letter u denotes the clock coordinate here, whereas it denoted the full state (x, c) there. Finally, the register clock is $c_i = u_i/\alpha$, so that the decoder $x_i > u_i/(2\alpha)$ used here coincides with the decoder $x_i - c_i/2 > 0$ used there.

Theorem D.1 (Rounded local computation in a convex energy)**.** *For every transition map F above, there is an explicit fixed energy $E_F : \mathbb{R}^{2m} \to \mathbb{R}$, independent of the initial bits and the horizon, with the following properties.*

1. *The energy is a sum of $O(m)$ factors of arity at most six; it is globally $C^{1,1}$, $3/4$-strongly convex, and $5/4$-smooth, with unique minimizer zero, and its step-one GD map is globally $1/4$-contractive.*

2. *A specified implementation rounds every operation to nearest with 37 significand bits and gradual underflow. Initialized by $x_i^0 = z_i^0 \in \{0,1\}$ and $u_i^0 = \alpha$, it satisfies, at every finite time,*

$$u_i^t = \alpha\gamma_t, \qquad \left|x_i^t/\gamma_t - (F^t(z^0))_i\right| \le 1/16.$$

After each raw update, each stored data value may be replaced by another represented value within $\gamma_{t+1}/64$ of it, although clock coordinates receive no such additional errors. Through horizon T, an exponent range of magnitude $O(T)$ suffices, using $O(\log(T+2))$ bits per exponent.

3. *Each update uses two message stages on the original gate-incidence graph, two dynamic features per register, and constant-size local modules. A finite family of ReLU modules implements the gradient formulas on every covered state, and its exact-real extension is globally 64τ-contractive, although off those states it need not be a gradient map.*

4. *Every covered rounded and perturbed update obeys $E_F(X^{t+1}) < 2\tau^2 E_F(X^t) < E_F(X^t)$.*

In words, one fixed strongly convex energy carries an arbitrary Boolean recurrence in the direction of its state while gradient descent shrinks the scale. Rounding to 37 bits with a horizon-dependent exponent range neither disturbs the clocks nor lets data errors accumulate.

D.1 The local factors and their purpose

The construction begins with the one-variable functions from which the factors are built. Fix the Boolean transition map F from Theorem D.1, let P_i be the set of its at most two distinct parent registers, let $S_i := \{i\} \cup P_i$ be its gate scope, and let $\Delta := \max_j |\{i : j \in S_i\}| \le 4$ be the maximum factor incidence. Repeated parent occurrences have been collapsed to unary truth tables, while self-inputs remain allowed. Write $x_i \in \mathbb{R}$ for a data coordinate, $u_i \in \mathbb{R}$ for its clock coordinate, and $X = (x, u) \in \mathbb{R}^{2m}$ for the full state. The fixed numerical parameters are

$$K := 25, \qquad \tau := 2^{-K}, \qquad p := K + 12 = 37, \qquad \alpha := 64, \qquad c_i := u_i/\alpha, \tag{29}$$

where p is the number of binary significand bits and c_i is the local signal scale. All scale changes in (29) are powers of two, and the rescaling α will make a clock feedback force smaller than a rounding interval without changing the number of local coordinates.

To suppress backward derivatives near a Boolean input, define the clipped quintic $P_0 : \mathbb{R} \to [0, 1]$ by

$$P_0(t) := \begin{cases} 0, & t \le 0, \\ 10t^3 - 15t^4 + 6t^5, & 0 < t < 1, \\ 1, & t \ge 1, \end{cases}$$

whose first two derivatives agree at the endpoints, so that $P_0 \in C^2$. On $(0, 1)$ they are $P_0'(t) = 30t^2(1-t)^2$ and $P_0''(t) = 60t(1-t)(1-2t)$, so $|P_0'| \le 15/8$ and $|P_0''| \le 6$. Define the bounded bump χ_0, the affine-neighborhood interpolant H, and the gate-input plateau R by

$$\begin{aligned} \chi_0(s) &:= P_0(2 - 8|s|), \\ H(q) &:= q\chi_0(q) + (q - \tfrac{1}{2})\chi_0(q - 1), \\ R(q) &:= P_0\big((q - 1/8)/(3/4)\big), \end{aligned}$$

where the absolute value introduces no singularity because χ_0 is constant near zero, and where the two summands of H have disjoint supports. For $z \in \{0, 1\}$ and $|q - z| \le 1/8$, these functions satisfy

$$H(q) = q - z/2, \qquad H'(q) = 1, \qquad H''(q) = 0, \qquad R(q) = z, \tag{30}$$

so that the affine part supplies a nonzero output derivative while the plateau makes every input derivative vanish.

Each gate uses the plateau values through its truth polynomial. For two parents with table values $f_{00}, f_{01}, f_{10}, f_{11} \in \{0, 1\}$, let g_i denote the multilinear extension

$$g_i(a, b) := f_{00} + (f_{10} - f_{00})a + (f_{01} - f_{00})b + (f_{11} - f_{10} - f_{01} + f_{00})ab,$$

with unary and constant gates using the corresponding affine and constant polynomials, and for a normalized data vector $q \in \mathbb{R}^m$ define the interpolated gate $G_i(q) := g_i(R(q_j) : j \in P_i)$. On $[0, 1]^2$ the truth polynomial is a convex combination of the table entries, so $0 \le G_i \le 1$. Its coefficient absolute sum is at most five because the displayed four coefficients have bounds $1, 1, 1, 2$, and this bound also covers XNOR, constants, and unary gates. On a Boolean $1/8$-neighborhood, $G_i = F_i(z)$ and all parent derivatives of G_i are zero.

Local clocks require a further cutoff to prevent unbounded derivatives when their scales are incomparable. Define the clock cutoff

$$\chi(r) := P_0(4r - 2)P_0(4 - 2r),$$

which is supported in $[1/2, 2]$ and equals one on $[3/4, 3/2]$, so that its derivatives vanish near the synchronized ratio one. The two factors have derivative supports $(1/2, 3/4)$ and $(3/2, 2)$, which are disjoint, so at most one product-rule term is active at any r and

$$|\chi'| \le 4 \cdot \frac{15}{8} = \frac{15}{2} < 12, \qquad |\chi''| \le 16 \cdot 6 + 16(15/8)^2 + 4 \cdot 6 < 180.$$

The constants 12 and 180 are convenient upper bounds carried through the calculation below, the second retaining all three product-rule terms.

For nonzero clocks indexed by S_i, define the local interaction

$$V_i(X) := c_i^2 H(x_i/c_i) G_i((x_j/c_j)_{j=1}^m) \prod_{j \in P_i \setminus \{i\}} \chi(c_i/c_j), \tag{31}$$

and set $V_i(X) := 0$ if one of those clocks vanishes. A self-parent contributes no clock cutoff because its ratio would be identically one. The proposed energy is

$$E_F(X) := \frac{1}{2} \sum_i x_i^2 + \frac{1-\tau}{2} \sum_i u_i^2 - \tau \sum_i V_i(X). \tag{32}$$

Each factor involves at most six physical coordinates, and the formulas are piecewise rational, including the explicitly defined values at zero clocks.

D.2 Global conditioning, including zero clocks

The local factors must remain controlled away from the computational states, because a small coefficient alone would not preserve convexity if their Hessians were unbounded. The estimate below supplies the global bound needed by Theorem D.1, and it is also what Lemma D.9 uses to compare energy values with norms.

Lemma D.2 (Regularity of a rounded-computation factor)**.** *Each factor V_i in (31), with its specified zero extension, is globally $C^{1,1}$ with gradient Lipschitz constant at most $C_0 := 2{,}000{,}000$, and its gradient is zero whenever an involved clock is zero. As a result, E_F in (32) is globally $3/4$-strongly convex and $5/4$-smooth, has unique minimizer zero, and its step-one GD map is globally $1/4$-contractive.*

Proof. Elementary product derivatives are bounded on the active support first; those bounds then extend across zero clocks, and the factors are finally summed by incidence. The support of χ and its derivatives implies that every nonzero product-derivative term has

$$\frac{1}{2} \le c_i/c_j \le 2 \qquad (j \in P_i), \tag{33}$$

a condition that a self-parent satisfies identically, so all involved clocks have the same sign and comparable magnitudes in an active term. An unbounded normalized data coordinate can occur in a derivative only multiplied by a derivative of its own H or R, whose compact derivative supports then bound that normalized coordinate.

The one-variable estimates come first. Since $|\chi_0'| \le 15$ and $|\chi_0''| \le 384$, disjointness of the two bumps gives

$$|H| \le 1, \qquad |H'| \le 13, \qquad |H''| \le 320.$$

For example, on the second bump $|q - 1/2| \le 3/4$, so the last two bounds follow from $1 + (3/4)15 < 13$ and $2 \cdot 15 + (3/4)384 = 318 < 320$, while the first bump has smaller bounds. The data support of a nonzero derivative of H lies in $|q| \le 5/4$, and the plateau R has $|R'| \le 5/2$, $|R''| \le 11$, and derivative support in $[1/8, 7/8]$.

To translate these estimates into physical-coordinate derivatives, put $s := c_i$ within this proof. For $H(x_i/s)$, the first-derivative numerators are H' and $-qH'$, and the second-derivative numerators are H'', $-(H' + qH'')$, and $2qH' + q^2 H''$. Since $|q| \le 5/4$ on the derivative support, $|H'| \le 13$, and $|H''| \le 320$, the first-derivative numerators are bounded by $(5/4) \cdot 13 < 17$ and the second-derivative numerators by $2(5/4)13 + (5/4)^2\, 320 < 533$. For $R(x_j/c_j)$, (33) bounds the rescaled

first derivatives by $5/|s|$ and the second derivatives by $52/s^2$, the largest clock numerator being at most

$$4\big(2(7/8)(5/2) + (7/8)^2 \cdot 11\big) < 52.$$

For $\chi(s/c_j)$, the rescaled first derivatives are at most $48/|s|$, and the largest second derivative is the parent-clock derivative, whose bound is

$$\frac{4(2 \cdot 2 \cdot 12 + 2^2 \cdot 180)}{s^2} = \frac{3072}{s^2},$$

while the own-clock and mixed-clock bounds are smaller. These calculations show that each elementary factor f of an active product obeys

$$|f| \le 1, \qquad |\partial_a f| \le 48/|s|, \qquad |\partial_{ab} f| \le 3072/s^2 \tag{34}$$

for all pairs of physical coordinates a, b, including $a = b$.

The product terms are counted next, without suppressing self-input overlap. Expand G_i into monomials in its at most two plateau values. Apart from a scalar coefficient, a resulting term of V_i is $s^2 A$, where A is a product of at most five factors: one H, two R factors, and two clock cutoffs. The derivative bounds in (34) give

$$|\partial_a A| \le 240/|s|, \qquad |\partial_{ab} A| \le \frac{5 \cdot 3072 + 20 \cdot 48^2}{s^2} = \frac{61440}{s^2},$$

and every Hessian entry of $s^2 A$ therefore has absolute value at most $2 + 4 \cdot 240 + 61440 = 62402$. The coefficient absolute sum is at most five, and a factor contains at most six coordinates, so the row-sum (Gershgorin) bound $\|W\|_{\mathrm{op}} \le \max_i \sum_j |W_{ij}|$ for a symmetric matrix W yields

$$\|\nabla^2 V_i\|_{\mathrm{op}} \le 6 \cdot 5 \cdot 62402 = 1{,}872{,}060 < C_0 \tag{35}$$

almost everywhere away from zero clocks. Because the product rule was applied in physical coordinates, a coordinate appearing in both H and a parent plateau has already contributed all its terms. Rescaling from c to u multiplies the Hessian on both sides by a diagonal contraction, since $\alpha \ge 1$, which cannot increase the operator norm.

These estimates also establish the zero extension. A factor satisfies $|V_i| \le c_i^2$, and its gradient is $O(|c_i|)$ uniformly in the data coordinates by the weighted first-derivative estimates just proved. If an involved clock vanishes while another remains nonzero, a ratio cutoff is zero throughout a neighborhood. If all involved clocks vanish, the displayed value and gradient estimates show that the value and gradient extend continuously by zero, which proves C^1 regularity. On a generic line segment, the nonzero-clock pieces are finitely many polynomial or rational pieces with matching gradients, so integrating (35) bounds the gradient difference by C_0 times the segment length. Continuity supplies the same bound at the finitely many zero-clock crossings, and exceptional segments follow by approximation. A segment lying in an involved zero-clock hyperplane has identically zero factor, so the extended gradient is globally Lipschitz.

Finally, let $h \in \mathbb{R}^{2m}$ be a test vector and write h_{S_i} for its entries at both coordinates of each register in S_i. Incidence counting gives

$$\left| h^T \nabla^2 \Big(\sum_i V_i \Big) h \right| \le C_0 \sum_i \|h_{S_i}\|^2 \le C_0 \Delta \|h\|^2,$$

and comparing (32) with $\|X\|^2/2$ gives

$$\|\nabla^2 E_F - I\|_{\mathrm{op}} \le \tau(1 + C_0\Delta) \le \frac{8{,}000{,}001}{33{,}554{,}432} < \frac{1}{4}.$$

Integrating these almost-everywhere Hessian bounds proves strong convexity, smoothness, and the contraction of $I - \nabla E_F$, and since every factor gradient vanishes at zero, zero is the unique minimizer. □

D.3 THE GRADIENT ON BOOLEAN TUBES AND THE NUMERICAL INVARIANT

The global estimate establishes a legitimate energy, but correct computation requires a stronger local identity: every input derivative must vanish on a neighborhood, while the output derivative restores a Boolean value. For $z \in \{0,1\}^m$ and $c > 0$, define the synchronized tube

$$\mathcal{T}(z,c) := \{X : u_i = \alpha c \text{ for every } i,\ |x_i/c - z_i| \le \delta \text{ for every } i\}, \qquad \delta := 1/16, \tag{36}$$

with data tolerance δ. The clock equality in its definition is maintained by rounding after each update, as Lemma D.5 will show, and the lemma below computes the gradient on this tube, which is the formula that every later lemma implements.

Lemma D.3 (Gradient on a Boolean tube)**.** *For $X \in \mathcal{T}(z,c)$, let $b := F(z)$ be the successor Boolean state. The energy gradients satisfy*

$$\partial_{x_i} E_F = x_i - \tau c b_i, \qquad \partial_{u_i} E_F = (1-\tau)u_i - (\tau/\alpha) b_i (x_i - z_i c), \tag{37}$$

and on an open neighborhood of each tube state the same identities hold with the Boolean labels retained and the common c on the right-hand sides replaced by the individual $c_i = u_i/\alpha$. These identities include all backward derivatives.

Proof. The flat input and clock regions identify the only surviving factor derivatives. Each normalized data argument lies strictly inside the $1/8$ neighborhood in (30), so $G_i = b_i$ and all its first derivatives vanish; every clock ratio is one, so its cutoff is one and its derivative is zero. The surviving part of the own factor is

$$c_i^2 H(x_i/c_i) b_i = b_i (c_i x_i - z_i c_i^2/2),$$

whose data derivative is $c_i b_i$ and whose c_i derivative is $b_i(x_i - z_i c_i)$, the latter acquiring the factor $1/\alpha$ when differentiated with respect to u_i. Self-inputs are included in this calculation, because their plateau derivatives vanish, and substitution into (32) gives (37). Strict containment in the flat regions also proves the neighborhood claim, with the varying individual c_i in the formulas. □

On an exactly encoded state the clock correction vanishes, which gives the exact orbit that the prediction theorem uses for its exact-arithmetic model.

Corollary D.4 (Exact execution by the same local energy)**.** *For the energy E_F in (32), ordinary gradient descent with step size one and exact arithmetic, initialized at $x_i^0 = z_i^0 \in \{0,1\}$ and $u_i^0 = \alpha$, satisfies*

$$x_i^t = \tau^t (F^t(z^0))_i, \qquad u_i^t = \alpha \tau^t \qquad (t \ge 0), \tag{38}$$

so the comparison $x_i^T > u_i^T/(2\alpha)$ returns $(F^T(z^0))_i$, with absolute margin $\tau^T/2$.

Proof. The point is that the clock correction in (37) vanishes at an exactly encoded Boolean state. Suppose $x_i = c z_i$ and $u_i = \alpha c$ for every i, where $c > 0$ and $z \in \{0,1\}^m$. This state belongs to $\mathcal{T}(z,c)$, and writing $b = F(z)$, Lemma D.3 gives

$$x_i^+ = x_i - (x_i - \tau c b_i) = \tau c b_i, \qquad u_i^+ = u_i - (1-\tau)u_i = \alpha \tau c,$$

so the next state is the exact encoding of $F(z)$ at scale τc. Induction from $c = 1$ proves (38). Subtracting the decoder threshold gives $\tau^T((F^T(z^0))_i - 1/2)$, which proves the readout and margin claims. □

Before analyzing the arithmetic, fix its model. Let RN round a real elementary result to a nearest radix-two number with p significand bits, with ties to even, and let $v := 2^{-p} = \tau/4096$ denote the unit roundoff. Until Lemma D.8, the exponent range is taken large enough for the specified horizon that underflow and overflow are excluded in the operations used. A nonzero rounded result then has relative error at most v, multiplication by a power of two is exact, and comparisons and ReLU preserve an already represented argument exactly.

The only other arithmetic fact needed is the exact subtraction of nearby floating-point numbers, commonly called Sterbenz's lemma (Higham, 2002b, Theorem 2.5): if represented positive numbers a, b satisfy $a/2 \le b \le 2a$, then $a - b$ is represented exactly, provided the exponent range contains it. To verify this, assume $a \ge b$ and choose an integer e with $2^e \le b < 2^{e+1}$. Both numbers are integer

multiples of 2^{e-p+1} because $a \le 2b < 2^{e+2}$, so their difference is a multiple of the same spacing and satisfies $0 \le a - b \le b < 2^{e+1}$, which means it has at most p significant bits.

The first use of these facts is the clock update, which must reproduce the exact power-of-two clock in spite of the small interaction force.

Lemma D.5 (Local clocks remain exactly synchronized)**.** *Suppose $X \in \mathcal{T}(z, c)$ is represented, $c = \tau^t$ for a nonnegative integer t, and $b = F(z)$. Compute, with every operation rounded,*

$$\begin{aligned} B_i &:= \mathrm{RN}((1-\tau)u_i), & r_i &:= \mathrm{RN}(x_i - z_i c), \\ h_i &:= (\tau/\alpha) b_i r_i, & g_{u_i} &:= \mathrm{RN}(B_i - h_i), & u_i^+ &:= \mathrm{RN}(u_i - g_{u_i}). \end{aligned} \tag{39}$$

Then $u_i^+ = \alpha\tau^{t+1}$ for every i.

Proof. The claim is that the small interaction force stays inside the rounding interval of the exactly represented quadratic gradient. Since the clock $u_i = \alpha\tau^t$ is a positive power of two and $1 - \tau$ has at most K significand bits, $B_i = (1-\tau)u_i$ without rounding error. This number is an interior point of the binade $[u_i/2, u_i)$, whose spacing is vu_i, so its distance to either nearest rounding boundary is $vu_i/2$.

The residual r_i is also exact: for $z_i = 0$ it subtracts zero, and for $z_i = 1$ the tube gives $x_i/c \in [15/16, 17/16]$, so Sterbenz's lemma applies. Bit selection and power-of-two scaling make h_i exact, and its magnitude obeys

$$|h_i| \le \frac{\tau}{\alpha} c\delta = \frac{\tau u_i \delta}{\alpha^2} = vu_i\delta < vu_i/2,$$

because $\alpha^2 = 4096$. Hence $g_{u_i} = B_i$ regardless of the nearest tie rule, and the final difference is the exactly represented number $u_i - B_i = \tau u_i$, which proves the clock identity without any communication between clocks. □

The data update must now restore the next bit at that exact clock scale, and its error estimate, which accounts for each arithmetic operation, includes the cancellation in subtracting the gradient; the lemma is the inductive step of the tube invariant.

Lemma D.6 (Data regeneration with persistent errors)**.** *Under the assumptions of Lemma D.5, compute*

$$g_i := \mathrm{RN}(x_i - \tau c b_i), \qquad x_i^{\mathrm{raw}} := \mathrm{RN}(x_i - g_i), \tag{40}$$

and after this computation allow any representable stored value x_i^+ with $|x_i^+ - x_i^{\mathrm{raw}}| \le \tau c/64$. Together with the clocks of Lemma D.5, the next state belongs to $\mathcal{T}(F(z), \tau c)$.

Proof. The clock lemma fixes the next denominator, so it remains to bound the data error relative to τc. If $b_i = 0$, the gradient is x_i and its subtraction gives zero, both without rounding error. If $b_i = 1$, put $a := x_i - \tau c$ and write $g_i = a + \eta$ with $|\eta| \le v|a|$. Rounding the final subtraction gives

$$\begin{aligned} |x_i^{\mathrm{raw}} - \tau c| &\le v|a| + v(\tau c + v|a|) \\ &\le v\big((1+v)(1+\delta+\tau) + \tau\big)c < 3vc, \end{aligned}$$

since $|x_i| \le (1+\delta)c$, and the permitted stored perturbation then gives

$$\left| \frac{x_i^+}{\tau c} - b_i \right| \le \frac{3v}{\tau} + \frac{1}{64} = \frac{3}{4096} + \frac{1}{64} < \frac{1}{16} = \delta.$$

The next tube property therefore holds for every coordinate and for every history-dependent choice of the permitted data errors. □

D.4 ReLU messages and finite exponent storage

The arithmetic identities above must be computable by the claimed local modules. On the invariant tube, ReLU expressions recover scaled Boolean values directly, so the implementation needs neither a real division oracle nor a discontinuous Boolean oracle. The following lemma builds those modules and bounds the Lipschitz constant of the complete block, establishing the theorem's contraction claim.

Lemma D.7 (Local ReLU realization and contraction)**.** *There is a fixed finite family of constant-size ReLU modules, selected by static gate and port types, that executes (37) on every represented synchronized tube state, with every operation rounded as specified in the preceding lemmas. One complete exact-real recurrent block extends to a globally Lipschitz map $\mathcal{M}_F : \mathbb{R}^{2m} \to \mathbb{R}^{2m}$ with*

$$\mathrm{Lip}(\mathcal{M}_F) \le 32\tau\sqrt{\Delta} \le 64\tau,$$

which equals $I - \nabla E_F$ on the synchronized tubes, although its behavior elsewhere need not be conservative. Each complete block uses a gather stage and a return stage on the factor graph and constant dynamic width.

Proof. Scaled Boolean gates are constructed first, the small clock residual is implemented next, and the complete real-valued map is bounded last. Let $\sigma(a) := \max(a, 0)$ denote ReLU, and for a positive local signal scale c define the restoration map

$$\mathcal{R}_c(x) := \sigma(2x - c/2) - \sigma(2x - 3c/2), \tag{41}$$

whose value is cz whenever $|x/c - z| \le 1/16$. To execute it in floating arithmetic, first compute $s := \mathrm{RN}(2x - c/2)$, then compute $\mathrm{RN}(s - c)$, apply the two ReLUs, and subtract their outputs. When $z = 0$, the first preactivation is in $[-5c/8, -3c/8]$ before rounding, so both ReLU outputs are zero. When $z = 1$, it is in $[11c/8, 13c/8]$ before rounding and its rounded value remains in $(c, 2c)$, so $s - c$ is exact by Sterbenz's lemma and the final difference is the represented number $s - (s - c) = c$. The arithmetic thus recovers 0 or c without approximation.

Use these restored amplitudes in a disjunctive normal form for the supplied truth table, where complement is $c - A$ for an amplitude A and conjunction and disjunction are minimum and maximum, both ReLU operations, for example $\min(A, B) = A - \sigma(A - B)$. A two-input truth table uses at most four terms of two literals, so the expression has fixed size, and all Boolean intermediates belong to $\{0, c\}$, which makes every operation exact. Let $J_i := cb_i$ denote the resulting gate-output amplitude and $Z_i := cz_i$ the receiver's restored old amplitude.

The clock gradient also uses the signed data residual. Compute $r_i := x_i - Z_i$ as in Lemma D.5, and select it using

$$Q_i := \sigma\big(r_i + (J_i - c)\big) - \sigma\big(-r_i + (J_i - c)\big), \tag{42}$$

where the parentheses prescribe the arithmetic order. If $J_i = c$, the inner difference is zero and the expression returns r_i, both without rounding error. If $J_i = 0$, both preactivations are strictly negative because $|r_i| \le \delta c$, so it returns zero. Hence $Q_i = b_i r_i$ exactly, and using τJ_i in the data gradient and $(\tau/\alpha)Q_i$ in the clock gradient reproduces the formulas (37) with the operation order already analyzed.

The same ReLU expressions define a map off the tube: receiver i uses $c_i = u_i/\alpha$ to restore its own and parent data, evaluates the same minimum/maximum truth expression, and forms Q_i by (42). Constant gates produce the output zero or c_i directly. In exact arithmetic, the full update is

$$(\mathcal{M}_F(X))_i = \tau\big(J_i,\ u_i + Q_i/\alpha\big), \tag{43}$$

a definition that makes sense for all real clock signs, although only its agreement with the energy gradient on the tube is needed for the computational claim.

To bound the extension globally, compare two physical states X and $\widetilde{X}$, let d_j be the Euclidean distance between their pairs at register j, and let $D_i := \max_{j \in S_i} d_j$ be the largest local input difference to gate i. ReLU's unit Lipschitz constant bounds the restoration difference by $4|\Delta x_j| + 2|\Delta c_i|$, and a complemented literal has bound $4|\Delta x_j| + 3|\Delta c_i|$. Minimum and maximum are nonexpansive in the sup norm of their arguments, so $|\Delta J_i| \le 7D_i$, and the residual has $|\Delta r_i| \le 7D_i$ as well. The selected residual therefore satisfies

$$|\Delta Q_i| \le 2|\Delta r_i| + 2|\Delta J_i - \Delta c_i| \le 30 D_i.$$

Since $\alpha \ge 1$, one output pair before the factor τ has squared difference at most $(7^2 + 31^2)D_i^2 < 32^2 D_i^2$. Summing over scopes gives $\sum_i D_i^2 \le \Delta \sum_j d_j^2$, which proves the displayed Lipschitz constant for the completed block.

Finally, each factor gathers the pairs in its scope and returns its local derivative contributions. On a tube every parent derivative is zero, and the nonzero own-coordinate formulas above are the factor

messages. The full piecewise rational factors supply the exact global gradient realization, while the ReLU branch extension supplies the conventional ReLU implementation on the computational states. Static type labels select from a fixed finite set of modules, and message and state widths are constant because gate arity and incidence are bounded. Contraction is asserted for the completed recurrent block, since individual communication stages need not contract. □

The preceding arguments excluded underflow in order to display the clock mechanism without extra notation, but a finite exponent range must also handle an adversary that stores a very small representable data error. The lemma below allows standard gradual underflow and closes that numerical boundary, which is what the resource statement of the theorem requires.

Lemma D.8 (A finite exponent range suffices)**.** *Fix an integer horizon $T \geq 0$, and use $p = 37$ significand bits, round to nearest, gradual underflow, and normal binary exponents ranging from $e_{\min} := -K(T+2) - 11$ to $e_{\max} := 10$. Then all conclusions of Lemmas D.5–D.7 remain valid through time T for every permitted data perturbation represented in this format, and each exponent can be stored in $O(\log(T+2))$ bits.*

Proof. Near zero the relative rounding bound is replaced by its absolute-error counterpart, and both invariants are then checked for sufficient slack. With $\gamma_T = \tau^T$ the final signal scale, the smallest positive subnormal is

$$q_{\min} := 2^{e_{\min}-p+1} = 2^{-K(T+2)-p-10} = 2^{-2K}\,\frac{v\gamma_T}{1024} \leq \frac{v\gamma_T}{1024},$$

where the exponent $K(T+2)$ in place of KT leaves a deliberate margin of $2^{2K} = 2^{50}$ beyond what the estimates below require. Every elementary result a in range satisfies

$$|\,\mathrm{RN}(a) - a| \leq v|a| + q_{\min}/2, \tag{44}$$

because for normal results the relative bound suffices, while for subnormal results rounding to the grid of spacing $q_{\min}$ gives the absolute term. Sterbenz subtraction remains exact with gradual underflow: if the difference is normal, the earlier significand argument applies, and if it is subnormal, both operands are integer multiples of $q_{\min}$, so the difference lies exactly on the subnormal grid.

The clock and restoration values are far above the smallest normal value, because every covered signal scale is at least $\gamma_T = 2^{-KT}$, whereas the smallest normal value is $2^{-K(T+2)-11}$. The baseline clock gradient $B_i = (1-\tau)u_i$ and its final subtraction therefore remain exactly represented, and only the very small feedback product h_i can underflow. Because that product is a power-of-two scaling of a represented residual, it is exact unless it enters the subnormal range, where its error is at most $q_{\min}/2$. The computed value $\widehat{h}_i$ therefore obeys

$$|\widehat{h}_i| \leq vu_i\delta + q_{\min}/2 < vu_i/2,$$

and the same nearest rounding interval as before yields $\mathrm{RN}(B_i - \widehat{h}_i) = B_i$, so the clocks still update exactly.

For the data update with $b_i = 1$, retain $a := x_i - \tau c$ and write $g_i = a + \eta$, where now (44) gives $|\eta| \leq v|a| + q_{\min}/2$. Applying it once more to the final subtraction bounds the raw data error by

$$\begin{aligned} |x_i^{\mathrm{raw}} - \tau c| &\leq v\big((1+v)(1+\delta+\tau)+\tau\big)c + (1+v/2)q_{\min} \\ &< 3vc. \end{aligned}$$

The last inequality follows from $q_{\min} \leq v\gamma_T/1024 \leq vc/1024$, valid because $c = \tau^t \geq \gamma_T$ for $t \leq T$, together with the fixed values of δ, τ, v, the first coefficient being less than $1.064v$. The zero-bit case still subtracts a represented number from itself exactly, so the same normalized data-error estimate closes.

The ReLU implementation also survives this finite format. Its restoration preactivations have magnitude comparable to c, so they are normal and the sign and exact-subtraction arguments are unchanged, and Boolean amplitudes and clock scales are normal as well. A selected residual may be subnormal, but ReLU, sign reversal, and subtraction of zero preserve it exactly, and its only potentially underflowing scaling is the feedback product already bounded. Overflow is excluded as well. All non-clock tube-module intermediates, namely restoration, Boolean amplitudes, and signed residual

selection, have magnitude at most $4c \le 4$, while clock-gradient intermediates have magnitude at most 65, since $u_i \le \alpha = 64$ and their corrections are smaller than one. Every intermediate is therefore below $65 < 2^7$, well inside the $e_{\max} = 10$ threshold. The allowed exponent interval has length $O(K(T+2)+p)$ with K, p fixed constants, which proves the bit cost. □

D.5 Descent and completion of the local theorem

The rounding invariant keeps the logic correct. To complete the theorem, we must also show that the perturbed execution decreases the same scalar objective. Because the clock coordinates dominate its Euclidean norm, this verification is independent of the number of registers. The next lemma supplies the decrease factor claimed in the theorem's last item.

Lemma D.9 (Energy decrease under the permitted data errors)**.** *If $X \in \mathcal{T}(z, c)$ and X^+ is any update covered by Lemmas D.6 and D.8, then*

$$E_F(X^+) < 2\tau^2 E_F(X) < E_F(X).$$

Proof. The norms are compared first, using the exact clock update, and global conditioning then compares the energy values. The old clock coordinates imply $\|X\|^2 \ge m\alpha^2 c^2$, while the next clock scale is τc and the next tube bounds every data coordinate by $(1+\delta)\tau c$, so

$$\|X^+\|^2 \le \tau^2 \left(1 + \frac{(1+\delta)^2}{\alpha^2}\right) \|X\|^2.$$

Lemma D.2, together with $E_F(0) = 0$ and $\nabla E_F(0) = 0$, gives $3\|X\|^2/8 \le E_F(X) \le 5\|X\|^2/8$, and combining these estimates shows

$$E_F(X^+) \le \frac{5}{3}\tau^2 \left(1 + \frac{(17/16)^2}{4096}\right) E_F(X) < 2\tau^2 E_F(X).$$

Since the positive clock makes $X \neq 0$, so that $E_F(X) > 0$, and since $2\tau^2 < 1$, both strict inequalities follow. □

Proof of Theorem D.1. The construction is (32), and the four items are collected from the lemmas in order. Lemma D.2 proves the arity, global regularity, conditioning, and minimizer claims of the first item. For the second, the initialization lies in $\mathcal{T}(z^0, 1)$, and Lemmas D.5 and D.6 inductively place every iterate in $\mathcal{T}(F^t(z^0), \gamma_t)$, including every allowed history-dependent data error. Lemma D.8 proves that this induction uses only the declared significand and exponent resources, and the linear comparison is correct because $\delta < 1/2$. Lemma D.7 proves the local ReLU realization and the contraction of a complete recurrent block asserted by the third item, its off-tube gradient distinction being part of that lemma's statement. Lemma D.9 gives the energy decrease of the fourth item, which completes the proof. □

The exponent range is a computational resource that grows with the horizon, and the construction uses ordinary round-to-nearest arithmetic with the specified order of operations. The allowed additional errors affect completed data registers only: independent clock faults could destroy the exactly represented baseline used by Lemma D.5, so no such stronger robustness statement follows from the theorem.

E The prediction problem and its complexity

Both arithmetic versions of Theorem 3.1 use a single local energy. Corollary D.4 establishes its exact Boolean orbit, and Theorem D.1 establishes its rounded implementation and readout invariant. Both are proved in Appendix D. This appendix specifies the common path architecture and the prediction language, then uses those two results to derive hardness and membership.

E.1 Fixed architecture and arithmetic

The hardness reduction needs a fixed machine whose local rule the compiler can implement, so we begin with that machine. Fix a deterministic Turing machine deciding the language TQBF of true fully quantified Boolean formulas, which is PSPACE-complete (Stockmeyer & Meyer, 1973), and choose a polynomial bound on its tape usage. Modify the machine so that after deciding it returns its head to the left boundary and enters an absorbing accepting or rejecting state, a modification that preserves polynomial space. An input path contains a finite-alphabet encoding of the input word, a canonical initial head, blank padding, and endpoint markers; its syntax is checkable in polynomial time, and invalid encodings are rejected.

The machine induces a fixed local rule on tape cells: the symbol of each cell records its tape symbol and, when present, the machine state at the head, and its next symbol depends only on that cell and its two neighbors. Extend this rule arbitrarily to invalid local configurations so that it is total, noting that only a fixed finite alphabet is used throughout.

Encode a cell symbol by a fixed number of Boolean bits, including a dedicated bit for the absorbing accepting state, and implement each next-state bit by a fixed circuit with fan-in at most two. Binary COPY trees reduce fan-out to at most two, including the broadcasts from a cell to the adjacent receiving circuits. Level the circuits, insert COPY registers until every wire joins consecutive levels, and include the feedback from the output level to the configuration registers. The resulting pipeline has a constant number D of levels, because the machine and its local rule are fixed.

Treat all circuit gates and configuration bits as persistent registers updated synchronously, with the following level convention. The configuration registers sit at level 0 and the remaining registers occupy levels $1, \dots, D-1$; every wire joins consecutive levels, and the feedback wires from level $D-1$ are the only inputs of the configuration registers. A register at level ℓ at time $t+\ell$ is then a fixed function of the level-0 block at time t. Writing $q(t)$ for the configuration-register block and $\mathcal{C}$ for the tape transition, substitution through the D levels gives

$$q(t+D) = \mathcal{C}(q(t)) \qquad (t \geq 0),$$

whatever values the other registers hold initially. Initialize $q(0)$ by the encoded tape and all other registers by zero; the subsequence $q(kD)$ then follows the machine, and the intermediate pipeline phases do not affect it. Each register is the output of one factor and an input to at most two factors, so the factor incidence satisfies $\Delta \leq 3$, which discharges the hypothesis $\Delta \leq 4$ of Appendix D.

For each tape length M, these fixed interior and boundary circuits define a Boolean register map F_M in which only the initial register values depend on the input word. Apply the local energy construction of Appendix D with $\tau = 2^{-25}$ and $\alpha = 64$, grouping all register pairs belonging to a tape cell into one vertex of the communication path $1 - 2 - \cdots - M$. The resulting energy, denoted by E_M, is independent of the input word and the horizon.

Each cell contains a constant number of gates and registers, and its internal circuit can branch and use feedback. Because every factor involves only constantly many neighboring cells, a constant number of nearest-neighbor message stages gathers its inputs and returns its derivative contributions, with left and right ports and endpoint markers determining the fixed routing rules. The constant-size ReLU modules of Lemma D.7 then evaluate the gradient on the computational states; their weights and the finite family of module types are independent of M, the input word, and the horizon. One complete recurrent message block implements one synchronous register update, which in exact arithmetic is one unit-step GD update, and a tape-machine transition takes D such blocks, so the horizon T below always counts complete GD blocks.

E.2 Exact and rounded evaluation

With the architecture fixed, the decision language can be stated in both arithmetic models. An input specifies the initial finite-alphabet tape and an integer $T \geq 0$ in binary; each data register is initialized by its encoded Boolean value and each clock by α. The only queried output is the fixed accepting-bit register at vertex 1, whose register index we denote by i_* and whose choice belongs to the architecture rather than to the input. In either arithmetic model, the input is accepted if and only if

$$x_{i_*}^T > u_{i_*}^T/(2\alpha). \tag{45}$$

For exact evaluation, every recurrent block is ordinary GD with step size one. Writing $z^t = F_M^t(z^0)$ for the Boolean orbit, Corollary D.4 proves

$$x_i^t = \tau^t z_i^t, \qquad u_i^t = \alpha\tau^t,$$

so that (45) returns $z_{i_*}^T$ for every input tape and every nonnegative horizon, including horizons between the simulated machine transitions.

For rounded evaluation, fix the operation order from Appendix D and use 37 significand bits, normal exponents from $-25(T+2)-11$ through 10, gradual underflow, and round to nearest with ties to even. In this decision language no stored data perturbation is applied, and from an exactly encoded Boolean state every rounded operation in the specified order returns a representable value. Each intermediate is a dyadic multiple of the current scale with at most $25 < 37$ significant bits, so the rounded trajectory coincides with the exact orbit bit for bit, and the $1/16$ tube of Theorem D.1 is needed only for the perturbation-robustness statement. Independently of this exactness, that theorem proves

$$u_i^t = \alpha\tau^t, \qquad |x_i^t/\tau^t - z_i^t| \le 1/16 \qquad (0 \le t \le T),$$

so that (45) again returns $z_{i_*}^T$, and the exact and rounded decision languages coincide on all valid inputs. The exponent bounds have $O(\log(T+2))$-bit descriptions; constant dynamic width counts coordinates, each rounded coordinate requires $O(\log(T+2))$ bits of storage, and one block uses $O(M)$ local arithmetic operations over the whole path.

E.3 Hardness and membership

Proof of Theorem 3.1. The proof reduces TQBF to the prediction language for hardness, simulates the Boolean network directly for membership, and then collects the analytic assertions from the compiler lemmas of Appendix D.

For hardness, let a TQBF instance have length n and choose a tape length M, polynomial in n, that accommodates the fixed decider. Its number of configurations on this tape is at most 2^{CM} for a fixed integer C, after increasing C to cover the head position, the control state, and the return to the left boundary. Since the decider halts, it has reached its absorbing decision by that time. Construct the canonical path encoding and set

$$T = D\,2^{CM}.$$

The path has polynomial length and T has $O(M)$ binary digits, so both are produced in polynomial time. The pipeline identity shows that the fixed output bit at time T is the TQBF decision, and the exact orbit identity and the rounded invariant both equate this bit with (45). This proves PSPACE-hardness for both arithmetic models under polynomial-time many-one reductions.

For membership, reject malformed inputs and simulate the Boolean register network synchronously. Since there are $O(M)$ registers, two Boolean buffers and a binary iteration counter use $O(M + \log(T+2))$ space, and both arithmetic models return the final Boolean output by the identities above. Because this argument applies to every valid input and horizon, it proves membership for each complete language, and it never materializes a fixed-point denominator of length proportional to T.

The analytic assertions come from the compiler. Lemma D.2 supplies global regularity, strong convexity, smoothness, and the unique minimizer, and grouping coordinates into path vertices preserves the Euclidean norm and with it the $1/4$ contraction bound for exact GD. Lemma D.7 shows that the fixed ReLU modules implement these exact GD steps on the encoded orbit and evaluate the gradient formulas on the covered rounded states. Their agreement with GD is required only there, because the globally defined ReLU extension need not be conservative elsewhere. Lemma D.9 proves strict energy decrease for the rounded implementation, and with it every assertion of the theorem is established. □

Composed with the reduction, the theorem rules out both fast uniform predictors and short verifiable certificates for the output bit, unless PSPACE collapses to P or to NP respectively.

Corollary E.1 (Consequences for prediction and certificates)**.** *A uniform algorithm deciding the language of Theorem 3.1 in time polynomial in the path length and* $\log(T+2)$ *implies* $\mathsf{P} = \mathsf{PSPACE}$. *If all accepting instances have polynomial-length certificates accepted by a single polynomial-time verifier that is sound on all rejecting instances, then* $\mathsf{NP} = \mathsf{PSPACE}$.

Proof. Composing the polynomial-time reduction with the proposed predictor or verifier places TQBF in P or in NP respectively, and the PSPACE-completeness of TQBF together with the usual reverse containments yields the claimed equalities. □

These statements concern a prescribed finite time given in binary, and polynomial or unary horizons admit direct polynomial-time simulation. Since the energy minimizer is the known origin, the classification supplies no hardness for approximating that minimizer to a fixed absolute accuracy. At time T the exact decoder margin is $\tau^T/2$ and the rounded margin is at least $7\tau^T/16$. These shrinking margins, together with an exponent range that must grow with T, are the reason the theorem gives no robustness guarantee for fixed absolute noise.

MATRIX REPRESENTATIONS AND PRESCRIBED GRADIENT DESCENT

F WHY CORRELATION COORDINATES MATTER

Section 4 opens with a one-edge comparison: two independent fair labels and one fair label copied with the opposite sign have the same means but cut the edge with probability $1/2$ and 1. This appendix extends that comparison to a weighted graph to distinguish enlarging a neural parameterization from enlarging the moments it represents. The statement concerns access to improvement directions at a specified state, and it neither gives a complexity lower bound nor obstructs escape from saddles by other means.

A *spin assignment* $\sigma \in \{-1, 1\}^n$ labels each of the n vertices by a sign, the two signs being the two sides of a cut. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric, nonzero, entrywise nonnegative weight matrix with zero diagonal, and let $W = \sum_{i<j} A_{ij} > 0$ be its total weight. The *weighted cut score* of an assignment counts the weight of the edges whose endpoints carry opposite signs,

$$c_A(\sigma) = \frac{1}{2}\sum_{i<j} A_{ij}(1 - \sigma_i\sigma_j).$$

A network that outputs only means describes a *product law*. For a mean vector $m \in [-1, 1]^n$, let μ_m be the law of n independent spins with $\mathbb{E}_{\mu_m}\sigma_i = m_i$, and write $C_A(m) = \mathbb{E}_{\mu_m} c_A(\sigma)$ for the expected cut score under this law, which we call the product extension of c_A. Correlations, by contrast, are carried by the second-moment matrix $\mathbb{E}\,\sigma\sigma^\top$ of a spin law, which always lies in the elliptope of unit-diagonal positive semidefinite matrices. The cut score extends linearly to that set,

$$\mathcal{E}_n = \{X \in \mathbb{R}^{n\times n} \mid X \succeq 0, \operatorname{diag} X = \mathbf{1}\}, \qquad L_A(X) = \frac{W}{2} - \frac{1}{4}\langle A, X\rangle,$$

where the Frobenius inner product counts both orientations of each edge, so that $L_A(\mathbb{E}\,\sigma\sigma^\top) = \mathbb{E}\, c_A(\sigma)$ for every spin law. Finally, the *moment map* $X_{\mathrm{prod}}(m) = \mathbb{E}_{\mu_m}\sigma\sigma^\top$ sends a mean vector to the second-moment matrix of its product law, so that a network of means reaches the elliptope only through X_{prod}. The proposition compares the gradient available through this map with the gradient available on the elliptope itself.

Proposition F.1 (First-order access to correlations)**.** *The product extension and its moment map satisfy*

$$C_A(m) = \frac{W}{2} - \frac{1}{4}m^\top Am, \qquad \nabla C_A(m) = -\frac{1}{2}Am, \tag{46}$$

$$X_{\mathrm{prod}}(m) = mm^\top + \operatorname{diag}(1 - m_i^2) \in \mathcal{E}_n, \qquad C_A = L_A \circ X_{\mathrm{prod}}. \tag{47}$$

At the unbiased center, $DX_{\mathrm{prod}}(0) = 0$. Hence every differentiable parameterization $\theta \mapsto m_\theta \in [-1, 1]^n$ with $m_{\theta_0} = 0$ obeys

$$\nabla_\theta[-C_A(m_\theta)]\big|_{\theta=\theta_0} = 0,$$

regardless of its parameter count. In contrast, the matrices $X_t = I - tA$ satisfy

$$X_t \in \mathcal{E}_n \quad (0 \le t \le \|A\|_{\mathrm{op}}^{-1}), \qquad L_A(X_t) = \frac{W}{2} + \frac{t}{4}\|A\|_F^2. \tag{48}$$

For $0 \le t \le 1/W$, X_t is the second-moment matrix of a zero-mean spin distribution. The product center is a saddle.

Proof. The product moments are computed directly, the moment map is differentiated at the center, and the improving elliptope direction is then exhibited together with a spin law that realizes it. Independence gives $\mathbb{E}_{\mu_m}\sigma_i\sigma_j = m_i m_j$ for $i \neq j$, proving (46), and since the diagonal moments equal one, the moment matrix is $mm^\top + \operatorname{diag}(1 - m_i^2)$. Both summands are positive semidefinite on the cube, which establishes (47). The diagonal of this matrix is constant and every off-diagonal entry is quadratic in m, so its differential vanishes at zero, and the chain rule then gives the parameter-gradient assertion.

For the stated spectral bound on t, every eigenvalue of $I - tA$ is nonnegative and its diagonal is one, and since $\langle A, I\rangle = 0$, substitution gives (48). To show that X_t is realizable by a spin law for sufficiently small t, set

$$h(\sigma) = \sum_{i<j} A_{ij}\sigma_i\sigma_j, \qquad \mu_t(\sigma) = 2^{-n}\big(1 - th(\sigma)\big).$$

The inequality $|h(\sigma)| \le W$ makes μ_t nonnegative when $t \le 1/W$, its total mass is one because $\mathbb{E}_{\mathrm{unif}} h = 0$, and orthogonality of spin monomials under the uniform law gives

$$\mathbb{E}_{\mu_t}\sigma_i = 0, \qquad \mathbb{E}_{\mu_t}\sigma_i\sigma_j = -tA_{ij} \quad (i \ne j),$$

so that its moment matrix is $I - tA$.

Finally, a nonzero symmetric matrix of trace zero has eigenvalues of both signs, so the Hessian $-A/2$ of C_A is indefinite at the interior point $m = 0$, which is the saddle assertion. □

The proposition says that at the uninformative center a mean-only network of any size has zero cut-loss gradient, while one correlation coordinate already exposes a first-order improvement. Perturbations and second-order methods can escape the product center when the relevant mean directions are accessible, and the proposition asserts nothing about other initializations. The product extension does retain the discrete optimum,

$$\max_{m\in[-1,1]^n} C_A(m) = \max_{\sigma\in\{-1,1\}^n} c_A(\sigma),$$

because every product expectation is an average of discrete scores and the product family includes every deterministic assignment; what correlation coordinates change is the set of available directions and rounding possibilities.

Two examples close the appendix: the first shows the gap between the means and the correlations of a single graph at its largest, and the second marks where the elliptope stops describing spin laws.

Example F.2 (Zero means with different cut values)**.** Suppose the positive-weight graph is connected and bipartite, and let $s \in \{-1, 1\}^n$ record its bipartition. The correlated law $(\delta_s + \delta_{-s})/2$ has zero means and cut score W, whereas the product law with the same zero means has score $W/2$.

In this case the SDP $\max_{X\in\mathcal{E}_n} L_A(X)$ has the unique optimum $ss^\top$. To see this, factor $X = VV^\top$ into unit rows and note that every edge contributes at most its weight, so attaining W forces $\langle v_i, v_j\rangle = -1$ and hence $v_i = -v_j$ across every positive edge. Connectivity then gives $v_i = s_i u$ for one unit vector u, so that $X = ss^\top$. Shared-Gaussian hyperplane rounding returns s or $-s$ almost surely, and both describe the same optimal cut. Since the graph itself is also solved by bipartite traversal, this example illustrates a representation difference without a computational separation.

Example F.3 (The boundary between a moment law and its relaxation)**.** For the unit-weight triangle, the matrix with diagonal one and every off-diagonal entry $-1/2$ is positive semidefinite, with eigenvalues $0, 3/2, 3/2$ and cut SDP score $9/4$, whereas every discrete cut has score at most 2. This matrix cannot be the second moment of any spin distribution, because

$$\sigma_1\sigma_2 + \sigma_1\sigma_3 + \sigma_2\sigma_3 \ge -1$$

holds for every spin assignment while the proposed matrix makes the expectation of the left side $-3/2$. Arbitrary elliptope points therefore require a rounding analysis of their own, which for cuts the Goemans–Williamson guarantee supplies (Goemans & Williamson, 1995), with the optimization error accounted for separately.

G Ordinary gradient descent with a polynomial spectral regularizer

This appendix proves Theorem 4.1, the first of the two theorems of Section 4, and its feasibility corollary, Corollary 4.2: ordinary constant-step gradient descent on the single loss (5), started at the scaled-identity initialization $V_0 = I_d/\sqrt{d}$ and run on a square factor, converges geometrically and returns a certificate read off the gradient. We first analyze a general convex matrix objective ϕ, then specialize to the log-sum-exp smoothing of the affine residuals.

The construction replaces the log-determinant barrier by a finite polynomial, so that every gradient is a product of matrices and no matrix solve is required. Although the polynomial remains finite at singular matrices, the specified initialization and step size keep the smallest eigenvalue of every normalized iterate above an explicit level r, and on that region a weighted log-determinant inequality (Lemma G.6) yields geometric convergence together with the gradient certificate. The final subsection records what goes wrong when one drops the scaled-identity initialization or the square factor, explaining the role of both requirements.

G.1 The polynomial loss and the general theorem

We begin with the general setting, because the log-sum-exp case in Theorem G.8 and the matrix-factorization results of Appendix H both instantiate one theorem. Write $\mathbb{S}^d$ for the real symmetric $d \times d$ matrices and let

$$\mathcal{D}_d = \{X \in \mathbb{S}^d \mid X \succeq 0,\ \operatorname{tr} X = 1\}, \qquad d \geq 2,$$

be the set of density matrices, that is, of positive semidefinite matrices of unit trace; all matrix inner products are Frobenius inner products $\langle A, H\rangle = \operatorname{tr}(AH)$. Suppose the objective ϕ is convex and twice continuously differentiable on an open convex neighborhood of $\mathcal{D}_d$, with a gradient bound $L > 0$ and a matrix smoothness constant $M \geq 0$ satisfying

$$\|\nabla\phi(X)\|_{\mathrm{op}} \leq L, \qquad 0 \leq \nabla^2\phi(X)[H,H] \leq M\|H\|_F^2 \quad (X \in \mathcal{D}_d,\ H \in \mathbb{S}^d). \tag{49}$$

Fix a barrier weight $\mu > 0$ and define the scale T, the eigenvalue level r, and the polynomial degree p by

$$T = L + \mu d, \qquad r = \frac{\mu}{4T}, \qquad p = \left\lceil \frac{1}{r} \log \frac{64T}{\mu} \right\rceil. \tag{50}$$

The polynomial regularizer, the geometric series h_p that appears in its gradient, and that gradient are

$$R_p(X) = \mu \sum_{j=1}^{p} \frac{\operatorname{tr}\big[(I-X)^j\big]}{j}, \qquad h_p(X) = \sum_{j=0}^{p-1} (I-X)^j, \qquad \nabla R_p(X) = -\mu h_p(X); \tag{51}$$

since $-\mu \log\det X = \mu \sum_{j\geq 1} \operatorname{tr}[(I-X)^j]/j$ whenever $0 \prec X \preceq I$, the regularizer is the degree-p truncation of the log-determinant barrier and $h_p(X)$ the corresponding truncation of X^{-1}. For every nonzero square factor $V \in \mathbb{R}^{d\times d}$, set

$$X(V) = \frac{VV^\top}{\|V\|_F^2}, \qquad \psi_p(X) = \phi(X) + R_p(X), \qquad \mathcal{L}_p(V) = \psi_p(X(V)). \tag{52}$$

Because the normalization is defined for every nonzero V and R_p is a polynomial, the loss $\mathcal{L}_p$ is smooth on all nonzero factors, singular ones included, and the work of the analysis is to show that the iterates nevertheless stay well conditioned.

Theorem G.1 (Ordinary GD with a finite polynomial regularizer; Theorem 4.1 restated)**.** *Under (49), start at the scaled-identity initialization $V_0 = I_d/\sqrt{d}$ and apply*

$$V_{t+1} = V_t - \eta \nabla \mathcal{L}_p(V_t), \qquad 0 < \eta \leq \frac{1}{64T + 4M}. \tag{53}$$

Write $s_t = \|V_t\|_F^2$ for the squared radius, $W_t = V_t/\sqrt{s_t}$ for the normalized factor, $G_t = \sqrt{s_t}\nabla\mathcal{L}_p(V_t)$ for the sphere gradient, and $h_{\min} = \eta/(1+4\eta L)$ for the smallest effective step. Every iterate satisfies

$$X(V_t) \succeq r I_d, \qquad 1 \leq s_t \leq 1 + 4\eta L. \tag{54}$$

The function ψ_p has a unique minimizer X_* on $\mathcal{D}_d$, with $X_* \succeq r I_d$. Moreover,

$$\mathcal{L}_p(V_t) - \psi_p(X_*) \leq 2L(1 - 4\mu h_{\min})^t, \tag{55}$$

$$\phi(X(V_t)) - \min_{X\in\mathcal{D}_d} \phi(X) \leq \mu d + \frac{\|G_t\|_F^2}{8\mu}. \tag{56}$$

Apart from evaluating $\nabla\phi$, the update and the certificate use only matrix products and scalar operations, and the factor stores d^2 real coordinates.

In words, from the scaled-identity start a fixed step of order $1/(T+M)$ never lets the normalized Gram matrix lose its conditioning, and the regularized loss contracts by the factor $1-4\mu h_{\min}$ at every step. The squared sphere gradient then bounds the gap of the original objective up to the additive μd that the regularizer costs.

The analysis works with the normalized factor. For a unit factor W, that is, one with $\|W\|_F = 1$, write $X = WW^\top$ and let A be the objective gradient, B the matrix gradient of ψ_p, λ_B its trace against X, and S the centered matrix gradient:

$$A = \nabla\phi(X), \quad B = A - \mu h_p(X), \quad \lambda_B = \langle B, X\rangle, \quad S = B - \lambda_B I_d. \tag{57}$$

Differentiating the normalization gives

$$G(W) = 2SW, \qquad \langle W, G(W)\rangle = 0, \qquad \nabla\mathcal{L}_p(V) = \frac{2}{\|V\|_F^2}\big(B - \langle B, X(V)\rangle I_d\big)V, \tag{58}$$

so that $G(W)$ is the gradient of $W \mapsto \psi_p(WW^\top)$ along the unit Frobenius sphere and the raw gradient is its rescaling. The polynomial h_p in this formula is evaluated by repeated squaring (see the arithmetic implementation below), whereas inverse matrices appear in what follows only as tools of the analysis, never in the algorithm.

The first lemma locates the target of the descent: the regularized minimizer exists, is unique, and lies inside the region $X \succeq rI_d$ to which the trajectory will be confined, so that the curvature comparison of Lemma G.4 can later be integrated along the segment from any iterate to it.

Lemma G.2 (An interior matrix minimizer)**.** *The minimizer X_* in Theorem G.1 exists, is unique, and satisfies $X_* \succeq rI_d$.*

Proof. Convexity and compactness give existence and uniqueness, after which a first-order comparison at a hypothetical null vector pushes the minimizer into the interior, where the stationarity condition bounds its smallest eigenvalue from below. For $0 \le x \le 1$, the geometric-series identity and Bernoulli's inequality $(1-x)^p \le (1+px)^{-1}$ give

$$xh_p(x) = 1 - (1-x)^p \in [0,1], \qquad h_p(x) \ge \frac{1}{x + 1/p}, \tag{59}$$

where the value at $x = 0$ is obtained by continuity; summing the first relation over the eigenvalues of X shows that $c(X) = \operatorname{tr}[Xh_p(X)] \le d$. The regularizer is convex on $\mathcal{D}_d$ because each $\operatorname{tr}[(I-X)^j]$ is a convex spectral function there, and its quadratic term $\mu\operatorname{tr}[(I-X)^2]/2$, present since $p \ge 1/r \ge 8$, makes the sum strictly convex; together with compactness of $\mathcal{D}_d$ this yields a unique minimizer.

If a minimizing X had a unit null vector u, the feasible direction $uu^\top - X$ would have objective derivative at most $2L + \mu d - \mu p$, which is negative because $p \ge 1/r = 4T/\mu$, so the minimizer is positive definite, and stationarity on the trace-one affine space makes its matrix gradient a scalar multiple of the identity. Testing this stationarity condition against an eigenvector of X_* for its smallest eigenvalue x_* gives

$$\mu h_p(x_*) \le 2L + \mu c(X_*) \le 2T,$$

and the second relation in (59) turns this into $x_* \ge \mu/(2T) - 1/p \ge r$, as claimed. □

G.2 A lower bound on the smallest eigenvalue along every trial step

The step size in the theorem is constant, so the descent estimate must hold not only at the iterates but at every point of the segment joining consecutive iterates. This subsection shows that the invariant $\lambda_{\min}(X) \ge r$ survives along the whole normalized trial path, which is what makes the derivative bounds of the next subsection uniform. The lemma is used twice: inductively in the proof of Theorem G.1, to carry the invariant from one iterate to the next, and inside Lemma G.5, to keep the trial path within the region where the loss has bounded derivatives.

Lemma G.3 (Invariant normalized trial path)**.** *If $\|W\|_F = 1$ and $WW^\top \succeq rI_d$, then*

$$W(h) = \frac{W - hG(W)}{\|W - hG(W)\|_F} \quad \textit{satisfies} \quad W(h)W(h)^\top \succeq rI_d \quad \left(0 \le h \le \frac{1}{64T}\right). \tag{60}$$

Proof. The idea is to compare the polynomial sphere gradient with the sphere gradient of the exact log-determinant loss, for which the Gram matrix of the trial point has a closed form. The difference between the two gradients then turns out to be too small to consume the eigenvalue margin that the exact barrier creates. The exact sphere gradient is

$$G_\infty(W) = 2[(A - \langle A, X\rangle I_d + \mu d I_d)W - \mu W^{-\top}].$$

Set $Q = (I - X)^p$ and $\delta = e^{-pr} \le \mu/(64T)$, so that $0 \preceq Q \preceq \delta I_d$ because $X \succeq r I_d$. Since $h_p(X) = X^{-1}(I - Q)$ and $\langle h_p(X), X\rangle = d - \operatorname{tr} Q$, the two gradients differ by

$$E = G(W) - G_\infty(W) = 2\mu[X^{-1}Q - \operatorname{tr}(Q) I_d]W.$$

Because Q commutes with X, $\|X^{-1}QW\|_F^2 = \operatorname{tr}(Q^2X^{-1}) \le \delta^2 \operatorname{tr} X^{-1} \le \delta^2 d/r$, while $\|\operatorname{tr}(Q)W\|_F = \operatorname{tr} Q \le d\delta$. Bounding the two terms of G_∞ separately by $4L + 2\mu d \le 4T$ and by $2\mu\|W^{-\top}\|_F \le 2\mu\sqrt{d/r} \le 4T$, where $\mu\sqrt{d/r} = 2\sqrt{\mu d T} \le 2T$ because $\mu d \le T$, we obtain

$$\|G_\infty(W)\|_F \le 8T, \qquad \|E\|_F \le 2\mu\delta(\sqrt{d/r} + d) \le 6T\delta, \qquad \|G(W)\|_F \le 9T, \tag{61}$$

the last bound because $\delta \le \mu/(64T) \le 1/128$.

Let $Y_\infty = W - hG_\infty(W)$ and $Y = W - hG(W)$ be the unnormalized trial points, with squared norms $q_\infty = \|Y_\infty\|_F^2$ and $q = \|Y\|_F^2$. With

$$D = I_d - 2h(A - \langle A, X\rangle I_d + \mu d I_d), \qquad a = 1 - 4hT,$$

we have $D \succeq a I_d$, and since $W^{-\top} = X^{-1}W$,

$$Y_\infty Y_\infty^\top = DXD + 4h\mu D + 4h^2\mu^2 X^{-1}.$$

Here $DXD \succeq rD^2 \succeq ra^2 I_d$ without any commutation between D and X, tangency $\langle W, G_\infty(W)\rangle = 0$ gives $q_\infty = 1 + h^2\|G_\infty(W)\|_F^2 \le 1 + 64h^2T^2$, and substituting $r = \mu/(4T)$ yields

$$Y_\infty Y_\infty^\top - r q_\infty I_d \succeq (2h\mu - 28h^2\mu T) I_d. \tag{62}$$

For $h \le 1/(64T)$, both $\|Y\|_F$ and $\|Y_\infty\|_F$ are below two, so (61) and tangency give

$$\|YY^\top - Y_\infty Y_\infty^\top\|_{\mathrm{op}} \le 4h\|E\|_F \le 24hT\delta,$$
$$|q - q_\infty| \le h^2\|E\|_F(\|G(W)\|_F + \|G_\infty(W)\|_F) \le 102h^2T^2\delta.$$

Passing from Y_∞ to Y therefore lowers the smallest eigenvalue of the left side of (62) by at most $24hT\delta + r\cdot 102h^2T^2\delta \le (24 + 102/64)hT\delta < 26hT\delta \le 26h\mu/64$, where we used $r \le 1/(4d) < 1$, $h \le 1/(64T)$, and $T\delta \le \mu/64$; for the same range of h the right side of (62) is at least $2h\mu - 28h\mu/64 = 100h\mu/64$. Hence $YY^\top \succeq rqI_d$, and dividing by $q \ge 1$ gives $W(h)W(h)^\top \succeq rI_d$, which is the claim. □

G.3 Descent and a weighted gradient inequality

With the invariant region in hand, we bound the curvature of the loss on it. Three lemmas follow: a two-sided comparison between the polynomial Hessian and the log-determinant Hessian, a uniform decrease along every trial step, and a gradient-domination inequality that converts the decrease into a geometric rate and a certificate. The comparison lets every curvature statement about R_p be transferred from the exact barrier $-\mu \log\det X$, whose Hessian $\mu \operatorname{tr}(X^{-1}HX^{-1}H)$ is explicit; its lower bound drives gradient domination and its upper bound the step-size condition.

Lemma G.4 (Spectral curvature comparison)**.** *For $X \in \mathcal{D}_d$ with $X \succeq rI_d$ and every symmetric H,*

$$\frac{\mu}{2} \operatorname{tr}(X^{-1}HX^{-1}H) \le \nabla^2 R_p(X)[H, H] \le \mu \operatorname{tr}(X^{-1}HX^{-1}H). \tag{63}$$

The upper bound also holds at every positive definite $X \in \mathcal{D}_d$.

Proof. The claim is spectral, so we establish the scalar version first and then pass to matrices through the divided-difference formula for the Hessian of a spectral function. For $r_p(x) = \mu\sum_{j=1}^p (1-x)^j/j$, differentiation gives

$$r_p''(x) = \frac{\mu}{x^2}[1 - (1-x)^p - px(1-x)^{p-1}],$$

which lies between zero and μ/x^2. If $x \geq r$, then $z = px \geq pr \geq \log(64T/\mu) \geq \log 128 > 4$, because $T/\mu \geq d \geq 2$, and since $(1-x)^p \leq e^{-z}$ and $px(1-x)^{p-1} \leq ez\,e^{-z}$,

$$(1-x)^p + px(1-x)^{p-1} \leq (1+ez)e^{-z} \leq \frac{1}{2},$$

the last step because $(1+ez)e^{-z}$ decreases for $z \geq 1$ and equals 0.22 at $z = 4$; hence $r_p''(x) \geq \mu/(2x^2)$ on $[r, 1]$. In an eigenbasis of X with eigenvalues x_i, the Hessian of $\operatorname{tr} r_p(X)$ has diagonal coefficients $r_p''(x_i)$ and off-diagonal coefficients $(r_p'(x_i) - r_p'(x_j))/(x_i - x_j)$, and integrating the scalar bounds on r_p'' between x_i and x_j compares each coefficient with the corresponding coefficient $\mu/(x_i x_j)$ of $-\mu \log\det X$. This proves (63), coincident eigenvalues being covered by continuity; the upper bound uses only $0 \leq r_p'' \leq \mu/x^2$, which holds on all of $(0, 1]$. □

The next lemma is the sufficient-decrease estimate. It is stated for a trial step of arbitrary size h up to the step-size threshold, rather than for the step the algorithm takes, because the iteration bound later applies it with the smallest effective step $h_{\min}$ at every iterate.

Lemma G.5 (Uniform trial decrease)**.** *Under the hypotheses of Lemma G.3,*

$$\psi_p(W(h)W(h)^\top) \leq \psi_p(WW^\top) - \frac{h}{2}\|G(W)\|_F^2 \quad \left(0 \leq h \leq \frac{1}{64T + 4M}\right). \tag{64}$$

Proof. We bound the first and second derivatives of the ambient loss on the invariant region and then differentiate along the trial path, which by Lemma G.3 stays inside that region. Consider $F(U) = \phi(UU^\top) + R_p(UU^\top)$ on ambient factors, which agrees with $\psi_p(UU^\top)$ at unit factors. At a unit factor with $UU^\top \succeq rI_d$,

$$\|\nabla F(U)\|_F \leq 2L + 2\mu\sqrt{d/r} \leq 6T.$$

The Hessian contribution from ϕ has norm at most $2L + 4M$, and for the regularizer the chain rule gives

$$\begin{aligned}\nabla^2(R_p \circ \mathrm{Gram})(U)[H,H] = {} & 2\langle -\mu h_p(X), HH^\top\rangle \\ & + \nabla^2 R_p(X)[UH^\top + HU^\top, UH^\top + HU^\top].\end{aligned}$$

The first term lies between $-2\mu\|H\|_F^2/r$ and zero because $0 \preceq h_p(X) \preceq X^{-1} \preceq I_d/r$, while the second is nonnegative and, by Lemma G.4 together with $\|X^{-1/2}U\|_{\mathrm{op}} = 1$ and $\|X^{-1/2}H\|_F \leq \|H\|_F/\sqrt{r}$, at most $4\mu\|H\|_F^2/r$. The regularizer therefore contributes a quadratic form between $-8T\|H\|_F^2$ and $16T\|H\|_F^2$, and altogether

$$\|\nabla^2 F(U)\|_{\mathrm{op}} \leq 18T + 4M.$$

Put $g = \|G(W)\|_F$. Since $G(W)$ is tangent to the sphere at W, the trial path is the unit-circle parametrization $W(h) = \cos\theta(h)\, W - \sin\theta(h)\, G(W)/g$ with $\theta(h) = \arctan(hg)$ when $g > 0$ and is constant when $g = 0$, so that $\|W'(h)\|_F \leq g$ and $\|W''(h)\|_F \leq 3g^2$ in either case. Because $h \leq 1/(64T + 4M) \leq 1/(64T)$, Lemma G.3 places every point of the path in the region where the derivative bounds above hold, and the second derivative of $F(W(h))$ is therefore at most $\|\nabla^2 F\|_{\mathrm{op}}\|W'\|_F^2 + \|\nabla F\|_F\|W''\|_F \leq (36T + 4M)g^2$, while its first derivative at zero is $-g^2$. Taylor's theorem gives $F(W(h)) \leq F(W) - hg^2 + \frac{h^2}{2}(36T + 4M)g^2 \leq F(W) - \frac{h}{2}g^2$, which is (64). □

A factor-Hessian variant. The proof used the matrix smoothness constant M only through the bound $2L + 4M$ on the Hessian of $\phi \circ \mathrm{Gram}$ in factor coordinates. If instead $|\nabla^2(\phi \circ \mathrm{Gram})(U)[H,H]| \leq (2L + M_F)\|H\|_F^2$ at every unit factor U with $UU^\top \succeq rI_d$, the same proof gives (64) for $0 \leq h \leq 1/(64T + M_F)$, because the second derivative of $F(W(h))$ is then at most $(36T + M_F)g^2$; Theorem G.1 then holds verbatim for $0 < \eta \leq 1/(64T + M_F)$, since its proof below uses M only through this lemma, and $M_F = 4M$ recovers the displayed version. The log-sum-exp objective of the next subsection is the case in which the variant matters, because its factor Hessian is smaller than the bound $4M$ obtained from its matrix Hessian by a factor d.

The inequality below is the engine of the gradient-domination step. It compares the Bregman divergence of $-\log\det$ with a norm weighted by $X^{-1/2}$ on one side only, and that asymmetric weighting is what allows the sphere gradient, which carries a factor $X^{1/2}$, to be paired with it by the Cauchy–Schwarz inequality.

Lemma G.6 (A weighted log-determinant inequality)**.** *For positive definite $X, Y \in \mathcal{D}_d$, define*

$$D_{-\log\det}(Y, X) = \operatorname{tr}(X^{-1}Y) - d - \log\det(X^{-1}Y).$$

Then

$$D_{-\log\det}(Y, X) \geq \|(Y - X)X^{-1/2}\|_F^2. \tag{65}$$

The constant one is sharp.

Proof. The proof reduces to a scalar inequality for the eigenvalues of the relative matrix $Q = X^{-1/2}YX^{-1/2}$, the two trace constraints entering as weights. Diagonalize Q with eigenvalues $q_i > 0$ and orthonormal eigenvectors u_i, and set $x_i = u_i^\top X u_i$, so that $x_i > 0$ and $\sum_i x_i = \sum_i x_i q_i = 1$ by the two trace constraints. In this notation the divergence is $\sum_i f(q_i)$ with $f(q) = q - 1 - \log q \geq 0$, while the squared norm in (65) is $\sum_i x_i(q_i - 1)^2$. If all $q_i = 1$, both sides vanish. Otherwise let $a = \min_i q_i < 1 < b = \max_i q_i$; averaging $(q_i - a)(b - q_i) \geq 0$ with the weights x_i gives

$$\sum_i x_i(q_i - 1)^2 \leq (b - 1)(1 - a),$$

whereas

$$\sum_i f(q_i) \geq f(a) + f(b) \geq (b - 1)(1 - a),$$

because $f \geq 0$ and $f(a) + f(b) - (b - 1)(1 - a) = ab - 1 - \log(ab) \geq 0$. Comparing the two displays proves the inequality, with no commutation assumption on X and Y. Equality holds for two distinct diagonal density matrices of size two whose diagonal entries are interchanged, so the constant one cannot be improved. □

Combining the curvature comparison with the weighted inequality gives the two estimates that drive the theorem: a gradient-domination inequality for the regularized loss, and a lower bound on $\lambda_{\min}(A)$, hence an upper bound on the Frank–Wolfe gap $\langle A, X\rangle - \lambda_{\min}(A)$ of Section 4.2, in terms of the same gradient norm.

Lemma G.7 (Gradient domination and an objective certificate)**.** *At every unit factor with $X = WW^\top \succeq rI_d$,*

$$\|G(W)\|_F^2 \geq 8\mu[\psi_p(X) - \psi_p(X_*)], \tag{66}$$

$$\langle A, X\rangle - \lambda_{\min}(A) \leq \mu d + \frac{\|G(W)\|_F^2}{8\mu}. \tag{67}$$

Proof. For the gradient inequality we integrate the lower Hessian bound along the segment from X to X_* and then apply the weighted inequality; for the certificate we complete a matrix square. The segment from X to X_* lies in $\{Y \succeq rI_d\}$ by Lemma G.2, so integrating the lower bound in (63) along it and then invoking Lemma G.6 gives

$$\begin{aligned}\psi_p(X) - \psi_p(X_*) &\leq \langle S, X - X_*\rangle - \frac{\mu}{2}\|(X_* - X)X^{-1/2}\|_F^2 \\ &\leq \frac{\|SX^{1/2}\|_F^2}{2\mu} = \frac{\|G(W)\|_F^2}{8\mu},\end{aligned}$$

where the second line is the weighted Cauchy–Schwarz inequality $\langle S, X - X_*\rangle \leq \|SX^{1/2}\|_F\|(X_* - X)X^{-1/2}\|_F$ followed by completing the scalar square, and $\|SX^{1/2}\|_F^2 = \operatorname{tr}(SXS) = \|G(W)\|_F^2/4$ because $G(W) = 2SW$; this proves (66).

For the certificate, set $\lambda_A = \langle A, X\rangle$ and $c = \operatorname{tr}[Xh_p(X)] \leq d$. Because $(I - X)^p \preceq e^{-pr}I_d \preceq I_d/2$, the series identity $h_p(X) = X^{-1}(I - (I - X)^p)$ gives $h_p(X) \succeq X^{-1}/2$. Substituting the decomposition

$$\lambda_A I_d - A = -S - \mu h_p(X) + \mu c I_d$$

into the matrix square

$$S + \frac{\mu}{2} X^{-1} + \frac{1}{2\mu} SXS \succeq 0$$

yields

$$A \succeq \left(\lambda_A - \mu d - \frac{\operatorname{tr}(SXS)}{2\mu} \right) I_d,$$

where the positive semidefinite matrix SXS was bounded by its trace times the identity. Since $\operatorname{tr}(SXS) = \|G(W)\|_F^2/4$, the smallest eigenvalue of A is at least $\lambda_A - \mu d - \|G(W)\|_F^2/(8\mu)$, which is (67). □

Proof of Theorem G.1. The lemmas are assembled by induction on t: the invariant $X_t \succeq r I_d$ propagates through Lemma G.3, the radius s_t is controlled by telescoping the descent estimate, and gradient domination converts descent into contraction. Homogeneity in (58) gives the identities

$$W_{t+1} = \frac{W_t - h_t G_t}{\|W_t - h_t G_t\|_F}, \qquad s_{t+1} = s_t + \frac{\eta^2}{s_t} \|G_t\|_F^2, \qquad h_t = \frac{\eta}{s_t},$$

so the normalized iterate moves along a trial path of Lemma G.3 with step h_t, while the squared radius grows by the squared step. Initially $s_0 = 1$ and $X_0 = I_d/d \succeq r I_d$ because $r \le 1/(4d)$; since s_t is nondecreasing, $h_t \le \eta \le 1/(64T)$, and the lower eigenvalue bound follows inductively from Lemma G.3. Lemma G.5 applied with $h = h_t$ gives $\psi_p(X_{t+1}) \le \psi_p(X_t) - \frac{h_t}{2} \|G_t\|_F^2$, and substituting $\|G_t\|_F^2 = (s_t/\eta^2)(s_{t+1} - s_t)$ turns this into

$$s_{t+1} - s_t \le 2\eta [\psi_p(X_t) - \psi_p(X_{t+1})].$$

The regularizer is minimized over $\mathcal{D}_d$ at I_d/d by convexity and unitary invariance, and the operator-norm bound on $\nabla\phi$ gives $|\phi(X) - \phi(Y)| \le L\|X - Y\|_* \le 2L$ for density matrices, where $\|\cdot\|_*$ is the nuclear norm; hence $\psi_p(X_0) - \psi_p(X_*) \le 2L$, and telescoping the radius inequality proves $s_t \le 1 + 4\eta L$, which is $h_t \ge h_{\min}$. Combining the descent estimate at step $h_t \ge h_{\min}$ with (66) yields

$$\psi_p(X_{t+1}) - \psi_p(X_*) \le (1 - 4\mu h_{\min})[\psi_p(X_t) - \psi_p(X_*)],$$

and iterating from $\psi_p(X_0) - \psi_p(X_*) \le 2L$ gives (55). Finally, convexity of ϕ bounds the unregularized gap by the Frank–Wolfe gap, $\phi(X) - \min_{\mathcal{D}_d} \phi \le \max_{Y \in \mathcal{D}_d} \langle A, X - Y \rangle = \langle A, X \rangle - \lambda_{\min}(A)$, and (67) turns this into (56). □

G.4 Affine feasibility and a finite stopping time

We now specialize the general theorem to the objective of Corollary 4.2, the log-sum-exp smoothing of finitely many affine residuals. This specialization uses two refinements: a step size based on the factor Hessian rather than on the matrix smoothness constant, and a dual interpretation of the certificate that bounds the smallest achievable maximum constraint violation. Consider signed affine residuals

$$a_i(X) = b_i + \langle B_i, X \rangle \quad (1 \le i \le m), \qquad a_0(X) = 0,$$

with symmetric coefficient matrices B_i, where the trivial residual has $B_0 = 0$ and $b_0 = 0$; let $q = m + 1 \ge 2$ count the residuals, and fix a positive bound $L \ge \max_i \|B_i\|_{\mathrm{op}}$ (if every B_i vanishes, the residuals are constant and feasibility is decided by inspection). Define the maximum constraint violation v, its minimum value γ over density matrices, the log-sum-exp smoothing ϕ_τ at temperature $\tau > 0$, and the softmax weights ω_i by

$$\begin{aligned} v(X) &= \max_{0 \le i \le m} a_i(X), \qquad \gamma = \min_{X \in \mathcal{D}_d} v(X), \\ \phi_\tau(X) &= \tau \log \sum_{i=0}^{m} e^{a_i(X)/\tau}, \qquad \omega_i(X) = \frac{e^{a_i(X)/\tau}}{\sum_{j=0}^{m} e^{a_j(X)/\tau}}. \end{aligned} \tag{68}$$

The zero residual ensures $\gamma \ge 0$, and compactness of $\mathcal{D}_d$ shows that the affine system is feasible if and only if $\gamma = 0$.

Theorem G.8 (Ordinary GD for bounded affine systems)**.** *Fix $0 < \varepsilon \le L$ and set*

$$\mu = \frac{\varepsilon}{4d}, \qquad \tau = \frac{\varepsilon}{4\log q}, \qquad \eta = \frac{1}{64(L+\mu d) + 4L^2/\tau}. \tag{69}$$

Use T, r, p from (50) and apply ordinary GD to $\mathcal{L}_p$ with $\phi = \phi_\tau$ from the scaled-identity initialization $V_0 = I_d/\sqrt{d}$. The invariants and the geometric convergence of Theorem G.1 hold, and at every iterate the quantity

$$\underline{\gamma} = \sum_i \omega_i(X) a_i(X) - \mu d - \frac{\|G\|_F^2}{8\mu} \le \gamma \le v(X) \tag{70}$$

brackets the smallest achievable maximum constraint violation, with

$$v(X) - \underline{\gamma} \le \tau \log q + \mu d + \frac{\|G\|_F^2}{8\mu}. \tag{71}$$

The stopping condition

$$\|V_t\|_F^2 \|\nabla \mathcal{L}_p(V_t)\|_F^2 = \|G_t\|_F^2 \le 4\mu\varepsilon \tag{72}$$

holds for every $t \ge N$, where

$$N = \left\lceil \frac{1}{4\mu h_{\min}} \max\left\{0, \log \frac{L}{\mu\varepsilon h_{\min}}\right\} \right\rceil, \qquad h_{\min} = \frac{\eta}{1+4\eta L}. \tag{73}$$

At termination, either $v(X) \le \varepsilon$ or $\underline{\gamma} > 0$ certifies infeasibility of the exact affine system, and the iteration bound is $\widetilde{O}\big(d(L/\varepsilon)^2 \log q\big)$ in exact real arithmetic.

Proof. The proof checks the hypotheses of Theorem G.1 in its factor-Hessian variant, reads the certificate of Lemma G.7 as a bound on the dual value, and converts the geometric rate into the stopping time. The gradient $A = \sum_i \omega_i B_i$ is a convex combination of the coefficient matrices, so $\|A\|_{\mathrm{op}} \le L$; a valid matrix smoothness constant is $M = dL^2/\tau$, but the factor curvature that the step size needs is smaller by a factor d. At every unit factor U,

$$\nabla^2(\phi_\tau \circ \mathrm{Gram})(U)[H,H] = 2\langle A, HH^\top\rangle + \frac{4}{\tau}\,\mathrm{Var}_{i\sim\omega}\langle B_i U, H\rangle.$$

The covariance operator is positive semidefinite with trace at most $4\sum_i \omega_i \|B_i U\|_F^2/\tau \le 4L^2/\tau$, so the factor Hessian satisfies $|\nabla^2(\phi_\tau \circ \mathrm{Gram})(U)[H,H]| \le (2L + 4L^2/\tau)\|H\|_F^2$. The factor-Hessian variant of Lemma G.5 with $M_F = 4L^2/\tau$ therefore gives Theorem G.1 for $0 < \eta \le 1/(64T + 4L^2/\tau)$, which is the step size in (69), together with the invariants and the geometric convergence.

For the certificate, (67) states that

$$A \succeq \alpha I_d, \qquad \alpha = \langle A, X\rangle - \mu d - \frac{\|G\|_F^2}{8\mu},$$

so for every density matrix Y,

$$v(Y) \ge \sum_i \omega_i(X) a_i(Y) \ge \sum_i \omega_i(X) b_i + \alpha = \underline{\gamma}.$$

Minimizing over Y gives $\underline{\gamma} \le \gamma$, and $\gamma \le v(X)$ holds by definition, which proves (70). For the gap, the entropy identity $\phi_\tau(X) = \sum_i \omega_i a_i(X) - \tau \sum_i \omega_i \log \omega_i$, the inequality $v(X) \le \phi_\tau(X)$, and the entropy bound $-\sum_i \omega_i \log \omega_i \le \log q$ combine into (71).

For the stopping time, let $\Delta_t = \psi_p(X_t) - \psi_p(X_*)$. Lemma G.5 applied at W_t with the trial step $h_{\min} \le \eta$, whether or not this is the step the algorithm takes, produces a point of value at least $\psi_p(X_*)$, so by (55) and $1 - x \le e^{-x}$,

$$\|G_t\|_F^2 \le \frac{2\Delta_t}{h_{\min}} \le \frac{4L}{h_{\min}} e^{-4\mu h_{\min} t}.$$

The majorant on the right decreases in t and drops below $4\mu\varepsilon$ once $t \ge \log(L/(\mu\varepsilon h_{\min}))/(4\mu h_{\min})$, so (72) holds for every $t \ge N$; the maximum in (73) is never active, because $h_{\min} \le 1/(64L)$ and

$\mu \le \varepsilon/8$ give $L/(\mu\varepsilon h_{\min}) \ge 512(L/\varepsilon)^2 \ge 512$. At the stated parameters, $64T = 64L + 16\varepsilon \le 80L$ and $4L^2/\tau = 16L^2 \log q/\varepsilon$, so

$$h_{\min}^{-1} = \eta^{-1} + 4L \le 84L + \frac{16L^2 \log q}{\varepsilon},$$

and $N = \widetilde{O}(d(L/\varepsilon)^2 \log q)$ follows. At stopping, the right side of (71) is at most $\tau \log q + \mu d + \varepsilon/2 = \varepsilon$; if $v(X) > \varepsilon$, the lower certificate $\underline{\gamma} \ge v(X) - \varepsilon$ is positive, and otherwise every affine residual is at most ε. □

Read as an algorithm, the theorem runs the gradient descent of Theorem G.1 at a temperature and a barrier weight that allot a quarter of the tolerance ε to the smoothing, a quarter to the regularizer, and half to the gradient test. It halts on that test with either an ε-feasible point or a dual witness of infeasibility. The constants $64T$, $8T$, and $9T$ in the step size and the gradient bounds were chosen to keep the proof short and are far from tight. Rerunning the nine trajectories of the accompanying script `check_polynomial_gd.py`, the stopping rule (72) fires at an iteration approximately 76–748 times smaller than the bound N of (73).

It remains to add the slack coordinate of Section 4.1. For a positive semidefinite variable $Z \in \mathbb{S}^n$ with a known bound $\operatorname{tr} Z \le R$, set $d = n + 1$ and read out $Z = RX_{1:n,1:n}$ from $X \in \mathcal{D}_d$; this slack-augmented trace-one embedding represents the entire trace-bounded cone, because

$$X = \operatorname{diag}(Z/R, 1 - \operatorname{tr} Z/R)$$

is a density matrix for every such Z. A residual $b_i + \langle C_i, Z\rangle$ becomes $a_i(X) = b_i + \langle B_i, X\rangle$ with $B_i = R \operatorname{diag}(C_i, 0)$, an equality contributes both of its signs, and positive rescalings of the residuals fix the units in which the final tolerance is measured. Applying Theorem G.8 to the embedded system proves Corollary 4.2: the output is positive semidefinite with trace at most R by construction, satisfies the remaining affine constraints to tolerance ε, and comes with the certificate $\underline{\gamma}$. Recovering a planted solution, rounding, or repairing the affine constraints exactly requires an argument specific to the application.

G.5 Arithmetic implementation

Since the theorem is stated for exact arithmetic, we record what one gradient step costs. The gradient (58) uses the geometric series $h_p(X)$ rather than the $1/j$-weighted value of $R_p(X)$, and the series is computed by repeated squaring: with $Q = I - X$, $P_j = Q^j$, and $S_j = \sum_{i=0}^{j-1} Q^i$, starting from $P_1 = Q$ and $S_1 = I_d$, the binary recursion

$$P_{2j} = P_j^2, \qquad S_{2j} = S_j + P_j S_j, \qquad P_{2j+1} = P_{2j} Q, \qquad S_{2j+1} = S_{2j} + P_{2j}$$

yields $h_p(X) = S_p$ after $O(\log p)$ dense matrix multiplications. The update and the stopping test use only this gradient and its norm, so the loss value is never evaluated, and for q dense affine residuals one step costs $O(d^3 \log p + qd^2)$ scalar arithmetic operations plus q exponentials. At the accuracy parameters of Theorem G.8,

$$p = O\left(\frac{dL}{\varepsilon} \log \frac{dL}{\varepsilon}\right),$$

and the factor occupies d^2 entries in addition to the storage for the input. The guarantee concerns exact real arithmetic, so a finite-precision implementation would need its own error analysis, although the recursion is at least well scaled: along the exact trajectory the spectrum of Q lies in $[0, 1 - r]$, so every power that appears has operator norm at most one.

G.6 Initialization and search-rank obstructions

The polynomial construction retains a square factor and the scaled-identity initialization, and the theorem charges for both: the factor stores d^2 coordinates, and the start is fixed rather than data dependent. The examples below show that neither requirement can be dropped within the same family of Gram losses, which is why the smaller search ranks of Section 4.2 come with a landscape statement rather than a convergence rate.

Proposition G.9 (A missed linear direction)**.** *Let a nonzero initial factor have a proper column space $\mathcal{S}$. There is a linear convex objective with a rank-one optimum such that ordinary GD on its normalized Gram loss remains suboptimal by one. The conclusion persists after adding any differentiable unitarily invariant spectral regularizer, whenever the iterates remain nonzero.*

Proof. The objective is chosen after the subspace, so that it is blind to everything the factor can represent. Choose a unit vector $u \perp \mathcal{S}$ and set $\phi(X) = -u^\top X u$, whose minimum on $\mathcal{D}_d$ is -1; on factors with columns in $\mathcal{S}$ its value is zero and its factor gradient vanishes. Because the matrix gradient of a unitarily invariant spectral regularizer commutes with X and preserves its support, each factor gradient has columns in $\mathcal{S}$, so induction keeps every iterate in that subspace and the original objective stays at zero. If the initial Gram matrix is $P_{\mathcal{S}}/\dim \mathcal{S}$, its nonzero eigenvalues coincide and the regularizer's factor gradient vanishes as well, so the trajectory is stationary. □

This proposition concerns a fixed proper initial subspace and permits an adversarial objective after the subspace has been chosen, so it leaves open guarantees for an informative, instance-dependent initialization.

Proposition G.10 (A strict local minimum for smoothed affine residuals)**.** *For $d = 2$ and rank-one factors on the unit sphere, a feasible system with three affine residuals can have a strict suboptimal local minimum of the factor loss built from its log-sum-exp smoothing. Every finite unitarily invariant spectral regularizer preserves this local minimum.*

Proof. The example makes the feasible point $e_2 e_2^\top$ the global minimizer of the smoothed loss while a strict local minimum sits at $e_1 e_1^\top$. Take $a > 0$, $\tau > 0$, and $c^2 > 2a\tau$, and set

$$a_0(X) = 0, \qquad a_\pm(X) = aX_{11} \pm cX_{12}.$$

Their log-sum-exp smoothing is

$$\phi_\tau(X) = \tau \log\bigl[1 + 2e^{aX_{11}/\tau} \cosh(cX_{12}/\tau)\bigr],$$

whose minimum is $\tau \log 3$ at $e_2 e_2^\top$ because $X_{11} \geq 0$ and $\cosh(cX_{12}/\tau) \geq 1$; since the maximum constraint violation is $v(X) = aX_{11} + |cX_{12}|$, the original system is feasible there. For $w(\theta) = (\cos\theta, \sin\theta)$, expansion at $\theta = 0$ gives

$$\phi_\tau(w(\theta)w(\theta)^\top) = \tau \log(1 + 2e^{a/\tau}) + \omega\left(\frac{c^2}{2\tau} - a\right)\theta^2 + O(\theta^4),$$

$$\omega = \frac{2e^{a/\tau}}{1 + 2e^{a/\tau}} > 0,$$

so $w = e_1$ is a strict local minimum with violation a and positive objective gap $\tau \log[(1 + 2e^{a/\tau})/3]$. All rank-one density matrices have the same spectrum, so every finite spectral regularizer is constant on the factor domain and cannot remove this minimum. The coefficient matrices have operator norm $(a + \sqrt{a^2 + c^2})/2$, and the choice $a = c = 1$ gives a fixed coefficient bound for every temperature $0 < \tau < 1/2$. □

Proposition G.11 (Suboptimal local minima at every proper search rank)**.** *For every $1 \leq k < d$, there is a convex smooth matrix objective with $\|\nabla\phi\|_{\mathrm{op}} \leq 5$ and $\nabla^2\phi[H, H] \leq 4\|H\|_F^2$ on $\mathcal{D}_d$ that has a rank-one global optimum and a rank-k factor local minimum on the unit Frobenius sphere with objective gap $2/k$. The local minimum is strict modulo right orthogonal transformations. Every finite convex unitarily invariant regularizer preserves a nonglobal local minimum at the same factor.*

Proof. The objective penalizes mass in a k-dimensional subspace linearly, penalizes it quadratically, and penalizes cross terms heavily, so that a factor spread uniformly over that subspace cannot move any mass out of it without first paying for the cross terms. Split $\mathbb{R}^d = \mathcal{S} \oplus \mathcal{S}^\perp$ with $\dim \mathcal{S} = k$ and write

$$X = \begin{pmatrix} B & Z \\ Z^\top & C \end{pmatrix}, \qquad \phi(X) = \frac{1}{k}\operatorname{tr} B + \|B\|_F^2 + 4\|Z\|_F^2.$$

This objective is nonnegative, equals zero at a rank-one projector in $\mathcal{S}^\perp$, and takes the value $2/k$ at $W_0 = [I_k \;\; 0]^\top/\sqrt{k}$. Writing a nearby unit factor as $W = [A^\top \;\; V^\top]^\top$ with $t = \|V\|_F^2$,

we have $B = AA^\top$, $Z = AV^\top$, and $\operatorname{tr} B = 1 - t$, and in a sufficiently small neighborhood $\sigma_{\min}(A)^2 > 3/(4k)$, whence

$$\phi(WW^\top) - \frac{2}{k} \geq -\frac{t}{k} + \frac{(1-t)^2 - 1}{k} + 4\sigma_{\min}(A)^2 t > 0 \qquad (t > 0).$$

When $t = 0$, equality requires $AA^\top = I_k/k$, which is the orbit of W_0 under right orthogonal transformations, so the local minimum is strict in the stated sense. The matrix Hessian is $2\|H_B\|_F^2 + 8\|H_Z\|_F^2 \leq 4\|H\|_F^2$, and since $\|B\|_{\mathrm{op}} \leq 1$ and $\|Z\|_{\mathrm{op}} \leq 1/2$ on $\mathcal{D}_d$, the block gradient has norm at most five.

A convex unitarily invariant regularizer is minimized over density matrices of rank at most k at the uniform rank-k spectrum, as one sees by averaging over all rotations within a k-dimensional space containing the support, so it preserves the local minimum at W_0. Replacing one column of W_0 by a vector of norm $1/\sqrt{k}$ in $\mathcal{S}^\perp$ produces a Gram matrix with the same spectrum and regularizer value but original objective $2(k-1)/k^2$, so the regularized local minimum has gap at least $2/k^2$, while the unregularized gap remains $2/k$. □

Shifted log-determinants and the finite polynomial regularizer above are among the regularizers covered by these statements, and two natural modifications fare no better. A log-determinant applied only to the rectangular right Gram matrix $W^\top W$ preserves the column space in Proposition G.9, and at search rank one its normalized right Gram matrix is the scalar one, so both obstructions persist.

H Matrix factorization and completion

This appendix proves Corollary H.2, together with its iteration bound, by applying the polynomial gradient-descent theorem of Appendix G to convex matrix losses with a known norm bound. The classical block positive semidefinite representation turns a nuclear-norm bound into a trace constraint (Vandenberghe & Boyd, 1996, Eq. (37)); see also Fazel et al. (2001, Lemma 1) and Jaggi & Sulovský (2010, Lemma 1 and Corollary 2). The latter paper already gives certified factorized optimization over this ball (Jaggi & Sulovský, 2010, Theorem 3). Here we specialize the prescribed GD construction to that established target. With this representation, ordinary GD from the scaled-identity initialization on a square factor returns rectangular output factors together with an additive certificate for their product.

For positive semidefinite sensing, adding one slack coordinate to the model matrix embeds the trace ball into the trace-one domain and gives the same conclusion for an output of the form $UU^\top$. The final subsection treats nuclear-norm completion, where an explicit correction of the observed entries converts approximate observations into exact ones.

H.1 A common accuracy prescription

Each corollary below is an instance of one prescription, recorded here once, that converts the regularized guarantee of Theorem G.1 into an accuracy bound for the original matrix objective. Throughout, $\mathcal{D}_d$ denotes the set of density matrices, that is, positive semidefinite $d \times d$ matrices of unit trace, introduced in Appendix G, with $d \geq 2$. The objective ϕ is convex and C^2 on an open convex neighborhood of $\mathcal{D}_d$, with a known positive gradient bound L and a known nonnegative smoothness bound M satisfying

$$\|\nabla\phi(X)\|_{\mathrm{op}} \leq L, \qquad 0 \leq \nabla^2\phi(X)[K,K] \leq M\|K\|_F^2 \quad (X \in \mathcal{D}_d, K \in \mathbb{S}^d). \tag{74}$$

For target accuracy $0 < \varepsilon \leq L$, set

$$\begin{aligned} \mu &= \frac{\varepsilon}{2d}, & T &= L + \mu d, & r &= \frac{\mu}{4T}, \\ p &= \left\lceil \frac{1}{r} \log \frac{64T}{\mu} \right\rceil, & \eta &= \frac{1}{64T + 4M}. \end{aligned} \tag{75}$$

For a nonzero raw factor $V \in \mathbb{R}^{d\times d}$, the normalized Gram matrix and the regularized loss are

$$X(V) = \frac{VV^\top}{\|V\|_F^2}, \qquad \mathcal{F}_p(V) = \phi(X(V)) + \mu \sum_{j=1}^{p} \frac{\operatorname{tr}[(I_d - X(V))^j]}{j}. \tag{76}$$

One fixed loss and one constant step size then serve each requested accuracy, and the lemma below records the resulting accuracy guarantee and iteration count, which every corollary in this appendix inherits.

Lemma H.1 (Original-objective accuracy)**.** *Under (74), start from the scaled-identity initialization $V_0 = I_d/\sqrt{d}$ and perform*

$$V_{t+1} = V_t - \eta \nabla \mathcal{F}_p(V_t), \tag{77}$$

checking the stopping condition below at every iterate, including iteration zero. At the first iterate satisfying

$$\|V_t\|_F^2 \|\nabla \mathcal{F}_p(V_t)\|_F^2 \le 4\mu\varepsilon, \tag{78}$$

the matrix $X_t = X(V_t)$ obeys

$$\phi(X_t) \le \min_{X \in \mathcal{D}_d} \phi(X) + \varepsilon. \tag{79}$$

Every iterate has $X_t \succeq rI_d$ and $1 \le \|V_t\|_F^2 \le 1 + 4\eta L$, and with $h_{\min} = \eta/(1+4\eta L)$ a sufficient iteration count is

$$N = \left\lceil \frac{1}{4\mu h_{\min}} \max\left\{0, \log \frac{L}{\mu\varepsilon h_{\min}}\right\} \right\rceil = \widetilde{O}\left(\frac{d(L+M)}{\varepsilon}\right). \tag{80}$$

Proof. The accuracy claim comes from the objective certificate of Theorem G.1, and the iteration count from its geometric rate together with a bound on the gradient by the regularized gap. The stated invariants are inherited directly, and the certificate (56), applied to the scaled gradient $G_t = \|V_t\|_F \nabla \mathcal{F}_p(V_t)$, yields

$$\phi(X_t) - \min_{X \in \mathcal{D}_d} \phi(X) \le \mu d + \frac{\|G_t\|_F^2}{8\mu},$$

whose right-hand side is at most $\varepsilon/2 + \varepsilon/2 = \varepsilon$ once (75) and (78) are inserted. For the iteration count, write Δ_t for the regularized matrix objective gap at iteration t. The same theorem gives the geometric rate on the left below, whereas the inequality on the right follows from the sufficient-decrease estimate by comparing its endpoint with the regularized minimum:

$$\Delta_t \le 2L \exp(-4\mu h_{\min} t), \qquad \|G_t\|_F^2 \le \frac{2\Delta_t}{h_{\min}}.$$

Together these force (78) by the iteration N in (80). Since $\varepsilon \le L$, the effective step satisfies

$$h_{\min}^{-1} = 64T + 4M + 4L \le 100L + 4M,$$

which gives the displayed asymptotic rate and completes the proof. □

Two boundary cases are handled separately: a zero gradient bound makes ϕ constant on $\mathcal{D}_d$, so there is nothing to optimize, whereas a requested accuracy larger than a positive bound L is simply replaced by the accuracy L. All iteration counts below are exact-real counts, and each step is additionally charged the operations needed to evaluate the matrix loss and its gradient.

H.2 Rectangular factorization over a nuclear-norm ball

Corollary H.2 (Matrix factorization over a nuclear-norm ball)**.** *Let f be convex and C^2 near the nuclear-norm ball of radius $R > 0$ in $\mathbb{R}^{a\times b}$, with known bounds $L_f > 0$ on $\|\nabla f\|_{\mathrm{op}}$ and $M_f \ge 0$ on its Hessian there. For $0 < \varepsilon \le RL_f$ and $d = a + b$, Theorem 4.1 applied to $\phi(X) = f(2RX_{12})$ returns, after finitely many exact-arithmetic steps, factors $U \in \mathbb{R}^{a\times d}$ and $H \in \mathbb{R}^{b\times d}$ with $\|UH^\top\|_* \le R$ and $f(UH^\top) \le \min_{\|Y\|_* \le R} f(Y) + \varepsilon$.*

The first application is the nuclear-norm ball of Corollary H.2. For a rectangular matrix Y, the nuclear norm $\|Y\|_*$ is the sum of its singular values, and it gives a convex budget on the product of two factors. The lemma below shows that a block representation reproduces that budget as the trace constraint of the normalized Gram readout, which is what allows the prescription above to be applied to a rectangular loss.

Lemma H.3 (A trace-one representation of a nuclear-norm ball)**.** *Let $a, b \geq 1$, $R > 0$, and $d = a + b$, and partition $X \in \mathcal{D}_d$ into blocks of sizes a and b. The readout*

$$Y = 2RX_{12}$$

maps $\mathcal{D}_d$ onto the nuclear-norm ball $\{Y \in \mathbb{R}^{a\times b} \mid \|Y\|_* \leq R\}$. If $X = WW^\top$ with $\|W\|_F = 1$, and the rows of W are partitioned as $W_1 \in \mathbb{R}^{a\times d}$ and $W_2 \in \mathbb{R}^{b\times d}$, then

$$U = \sqrt{2R}\, W_1, \qquad H = \sqrt{2R}\, W_2, \qquad Y = UH^\top, \qquad \|U\|_F^2 + \|H\|_F^2 = 2R. \tag{81}$$

Proof. The image lies in the ball because of a product inequality for the factors, and every point of the ball has a preimage built from a singular-value decomposition. For the displayed factors, writing $UH^\top$ as a sum of outer products of corresponding columns and applying Cauchy–Schwarz gives

$$\|UH^\top\|_* \leq \|U\|_F \|H\|_F \leq \frac{\|U\|_F^2 + \|H\|_F^2}{2} = R,$$

so the image is contained in the ball.

Conversely, given Y with $\|Y\|_* \leq R$, take a thin singular-value decomposition $Y = P\Sigma Q^\top$ with $P^\top P = Q^\top Q = I$. The block matrix

$$Z_0 = \begin{pmatrix} P\Sigma P^\top & Y \\ Y^\top & Q\Sigma Q^\top \end{pmatrix} = \begin{bmatrix} P \\ Q \end{bmatrix} \Sigma \begin{bmatrix} P \\ Q \end{bmatrix}^\top$$

is then positive semidefinite with trace $\operatorname{tr}(P\Sigma P^\top) + \operatorname{tr}(Q\Sigma Q^\top) = 2\|Y\|_*$. Adding a positive semidefinite matrix supported on its first diagonal block with trace $2(R - \|Y\|_*)$ produces a matrix Z of trace $2R$ whose off-diagonal block is still Y, and its normalization $X = Z/(2R)$ is the required density matrix, which shows that the readout is onto the ball. □

The decomposition appears only in this surjectivity argument, since the GD update itself acts on the raw factor and never forms one. Applying Lemma H.1 to the composed loss gives Corollary H.2 in quantitative form.

Corollary H.4 (Rectangular factorization with an accuracy certificate)**.** *Let f be convex and C^2 on an open convex neighborhood of the nuclear-norm ball of radius $R > 0$ in $\mathbb{R}^{a\times b}$. Suppose that known constants $L_f > 0$ and $M_f \geq 0$ satisfy*

$$\|\nabla f(Y)\|_{\mathrm{op}} \leq L_f, \qquad 0 \leq \nabla^2 f(Y)[K, K] \leq M_f \|K\|_F^2$$

for every Y in that ball and every rectangular perturbation K. Set

$$d = a + b, \qquad L = RL_f, \qquad M = 2R^2 M_f, \qquad \phi(X) = f(2RX_{12}). \tag{82}$$

For $0 < \varepsilon \leq L$, apply (75)–(77) to this ϕ. At the stopping condition (78), the factors in (81), formed from $W = V/\|V\|_F$, satisfy

$$\|UH^\top\|_* \leq R, \qquad f(UH^\top) \leq \min_{\|Y\|_* \leq R} f(Y) + \varepsilon. \tag{83}$$

The raw variable has d^2 entries, both output factors have d columns, and the iteration bound is (80).

Proof. We verify the two derivative bounds (74) for the composed objective and then let Lemma H.3 identify the comparison class. Since the readout $X \mapsto 2RX_{12}$ is linear, ϕ is convex and C^2 on a neighborhood of $\mathcal{D}_d$, with matrix gradient

$$\nabla\phi(X) = \begin{pmatrix} 0 & R\nabla f(2RX_{12}) \\ R\nabla f(2RX_{12})^\top & 0 \end{pmatrix},$$

a symmetric block matrix whose operator norm equals $R\|\nabla f(2RX_{12})\|_{\mathrm{op}} \leq L$. For a symmetric perturbation K, the chain rule gives

$$\begin{aligned} \nabla^2\phi(X)[K, K] &= \nabla^2 f(2RX_{12})[2RK_{12}, 2RK_{12}] \\ &\leq 4R^2 M_f \|K_{12}\|_F^2 \leq 2R^2 M_f \|K\|_F^2, \end{aligned}$$

where the last step uses $2\|K_{12}\|_F^2 \leq \|K\|_F^2$ for symmetric K. Hence (74) holds with the constants (82), so Lemma H.1 gives $\phi(X_t) \leq \min_{\mathcal{D}_d} \phi + \varepsilon$ at the stopping condition, and Lemma H.3 identifies $\min_{\mathcal{D}_d} \phi$ with $\min_{\|Y\|_* \leq R} f(Y)$ and the readout of X_t with $UH^\top$, which yields (83). □

The Euclidean update acts on the raw square factor V through the fixed loss (76), and the matrices U, H are read out from it only at the end. Any smooth convex fitting loss satisfying the displayed bounds is admissible. The guarantee is optimization over the nuclear-norm ball with a square factor of search rank d, although it says nothing about a prescribed rank constraint or about GD applied directly to $f(UH^\top)$. The smallest case shows how the representation trades a norm bound for a direction.

Example H.5 (A scalar product with an exact norm bound)**.** Take $a = b = 1$, so that the matrix variable is a scalar y whose nuclear-norm ball is the interval $[-R, R]$. The density matrix

$$X = \frac{1}{2}\begin{pmatrix} 1 & y/R \\ y/R & 1 \end{pmatrix}$$

is positive semidefinite if and only if $|y| \le R$, and its off-diagonal readout is $2RX_{12} = y$. One pair of output row factors is

$$U = \sqrt{R/2}\big(\sqrt{1+y/R}, \sqrt{1-y/R}\big), \qquad H = \sqrt{R/2}\big(\sqrt{1+y/R}, -\sqrt{1-y/R}\big),$$

which satisfy $UH^\top = y$ and $\|U\|^2 + \|H\|^2 = 2R$, so the factor directions encode every scalar in the interval while the combined factor norm stays fixed. Because the density matrix is positive definite in the open interval and singular at the endpoints, the GD theorem returns an objective-accurate interior output even when the comparison optimum is an endpoint.

H.3 Positive semidefinite matrix sensing

When the target is itself positive semidefinite, the block representation is unnecessary: the model matrix is already a Gram matrix, and a trace bound is represented by adding one slack coordinate. The sensing map may be arbitrary because the corollary compares fitting objectives over the entire bounded positive semidefinite domain rather than recovering a particular matrix.

Corollary H.6 (Positive semidefinite sensing)**.** *Let $n \ge 1$, let $\mathcal{A}\colon \mathbb{S}^n \to \mathbb{R}^m$ be a linear sensing map, let $y \in \mathbb{R}^m$ be the measurement vector, let $R > 0$ be the trace bound, and let $\kappa > 0$ be a known constant with*

$$\|\mathcal{A}(K)\|_2 \le \kappa \|K\|_F \quad (K \in \mathbb{S}^n).$$

Define

$$\begin{aligned} \mathrm{OPT}_R = \min_{Z \succeq 0,\ \mathrm{tr}\, Z \le R} \frac{1}{2}\|\mathcal{A}(Z) - y\|_2^2, \qquad & d = n+1, \\ L = R\kappa(R\kappa + \|y\|_2), \qquad & M = R^2\kappa^2. \end{aligned} \tag{84}$$

For $0 < \varepsilon \le L$, run the prescription of Lemma H.1 with

$$\phi(X) = \frac{1}{2}\|\mathcal{A}(RX_{1:n,1:n}) - y\|_2^2$$

and, at its stopping condition, read out the factor

$$U = \sqrt{R}\,\frac{V_{1:n,:}}{\|V\|_F} \in \mathbb{R}^{n \times (n+1)}. \tag{85}$$

Then

$$\|U\|_F^2 \le R, \qquad \frac{1}{2}\|\mathcal{A}(UU^\top) - y\|_2^2 \le \mathrm{OPT}_R + \varepsilon, \tag{86}$$

and the iteration count is (80) with the constants in (84).

Proof. The argument has two parts: the readout $Z = RX_{1:n,1:n}$ must cover the trace ball, and ϕ must satisfy the derivative bounds (74) with the stated constants. For the first part, every positive semidefinite Z with $\mathrm{tr}\, Z \le R$ is the readout of the density matrix

$$X = \mathrm{diag}\big(Z/R, 1 - \mathrm{tr}\, Z/R\big),$$

obtained by adding one slack coordinate, so the image of $\mathcal{D}_{n+1}$ is the whole trace ball. At the stopping iterate, the output (85) satisfies $RX(V_t)_{1:n,1:n} = UU^\top$.

For the derivative bounds, let $\mathcal{A}^*$ denote the adjoint of $\mathcal{A}$ for the Frobenius and Euclidean inner products, whose operator norm from Euclidean to Frobenius space is also at most κ. The nonzero block of $\nabla\phi(X)$ is

$$R\mathcal{A}^*(\mathcal{A}(RX_{1:n,1:n}) - y),$$

and since a positive semidefinite trace-one matrix has Frobenius norm at most one, this block has operator norm at most $R\kappa(R\kappa + \|y\|_2) = L$. For a symmetric perturbation K, the second derivative is

$$\nabla^2\phi(X)[K,K] = R^2\|\mathcal{A}(K_{1:n,1:n})\|_2^2 \le R^2\kappa^2\|K\|_F^2,$$

so ϕ is a convex quadratic satisfying (74) with $M = R^2\kappa^2$, and Lemma H.1 gives (86). □

Each iteration evaluates $\mathcal{A}$ and $\mathcal{A}^*$ once and spends $O(d^3 \log p)$ additional dense arithmetic operations on the polynomial gradient and the factor update, where $p = O((dL/\varepsilon)\log(dL/\varepsilon))$ for $0 < \varepsilon \le L$; no matrix inverse, eigendecomposition, or projection is required. The zero sensing map, for which the fitting objective is constant, needs no iteration at all.

The guarantee is an optimization statement that uses no restricted-isometry or sampling assumption, and for the same reason it makes no recovery claim. The zero sensing map, for instance, makes every matrix observationally indistinguishable, so recovering a planted low-rank matrix additionally requires that the measurements identify it within the model class. Neither a prescribed-rank factor nor a guarantee for the unregularized factor loss follows, and related factorization and recovery results address these issues under their respective assumptions (Bhojanapalli et al., 2016; Zhang et al., 2023; Ge et al., 2017).

H.4 Exact observed entries and approximate minimum nuclear norm

The last application replaces a fitting loss by hard constraints. In the nuclear-norm formulation of matrix completion (Candès & Recht, 2009), the affine observation constraints have a simple repair, since correcting an observed entry costs at most the magnitude of that correction in nuclear norm. We combine this repair with feasibility tests from Theorem G.8, each of which is a separate constant-step GD run on one fixed loss, inside an outer bisection over the budget.

Let $a, b \ge 1$, let $\Omega \subseteq \{1,\dots,a\} \times \{1,\dots,b\}$ be a set of s distinct observed positions with specified values $b_{ij} \in \mathbb{R}$, and define the optimal value $\mathrm{OPT}_{\mathrm{comp}}$ and the total observed mass B_0 by

$$\begin{aligned} \mathrm{OPT}_{\mathrm{comp}} &= \min\{\|Y\|_* \mid Y \in \mathbb{R}^{a\times b},\ Y_{ij} = b_{ij} \text{ for } (i,j) \in \Omega\}, \\ B_0 &= \sum_{(i,j)\in\Omega} |b_{ij}|. \end{aligned} \tag{87}$$

The zero-filled observation matrix $Y^{(0)}$ is feasible with $\|Y^{(0)}\|_* \le B_0$, by decomposition into single-entry matrices, and because a bounded nuclear-norm sublevel set is compact the minimum is attained.

Corollary H.7 (Completion with exact observed entries)**.** *For any accuracy $\varepsilon > 0$, there is a finite procedure that returns a matrix $\widehat{Y}$ satisfying*

$$\widehat{Y}_{ij} = b_{ij} \quad ((i,j) \in \Omega), \qquad \|\widehat{Y}\|_* \le \mathrm{OPT}_{\mathrm{comp}} + \varepsilon. \tag{88}$$

If $B_0 = 0$ or $\varepsilon \ge B_0$, the zero-filled matrix suffices. Otherwise the procedure uses at most

$$J = \left\lceil \log_2 \frac{2B_0}{\varepsilon} \right\rceil \tag{89}$$

feasibility tests. Each test applies Theorem G.8 with the known inputs

$$d = a + b + 1, \qquad q = 2s + 2, \qquad L = 2B_0, \qquad \delta = \frac{\varepsilon}{4(s + 1/2)} \tag{90}$$

and target residual accuracy δ. Each test uses $\widetilde{O}(d(L/\delta)^2 \log q)$ ordinary GD steps on a square factor with d^2 entries, while entry repair and the outer interval updates use scalar arithmetic and require no singular-value decomposition.

Proof. We construct a feasibility test from Theorem G.8, bound the nuclear-norm cost of repairing the observed entries, and then run a bisection that maintains an interval provably containing the optimum. Assume $0 < \varepsilon < B_0$; since $B_0 > 0$ forces at least one nonzero observed entry, $s \geq 1$. For a proposed budget $t \in [0, B_0]$, consider the system

$$Z = \begin{pmatrix} P & Y \\ Y^\top & Q \end{pmatrix} \succeq 0, \qquad \operatorname{tr} Z \leq 2t, \qquad Y_{ij} = b_{ij} \quad ((i,j) \in \Omega),$$

in a symmetric block matrix Z with blocks of sizes a and b. This system is feasible if and only if a completion of nuclear norm at most t exists. The block construction in Lemma H.3 produces such a Z from any completion, whereas a Gram factorization of any feasible Z and the product inequality in that lemma give $\|Y\|_* \leq \operatorname{tr} Z/2 \leq t$.

To pose this as a bounded affine system, use the fixed trace bound $R_0 = 2B_0$ for every test, with density dimension $d = a + b + 1$ and readout $Z = R_0 X_{1:a+b,1:a+b}$; because $t \leq B_0$, the bound excludes no feasible feasibility test. The affine residuals are both signs of $Y_{ij} - b_{ij}$ for each observed position, the single residual $\operatorname{tr} Z - 2t$, and the zero residual, which makes $q = 2s + 2$ residuals in all. An entry residual has a symmetric coefficient matrix with two off-diagonal entries $R_0/2$ and hence operator norm $R_0/2$, while the trace residual has coefficient $R_0 \operatorname{diag}(I_{a+b}, 0)$ of norm R_0. The parameters (90) are therefore valid coefficient bounds with the correct residual count, and $0 < \delta < L$ holds because $\varepsilon < B_0$.

At termination, Theorem G.8 either certifies that this exact system is infeasible or returns a positive semidefinite Z with

$$|Y_{ij} - b_{ij}| \leq \delta \quad ((i,j) \in \Omega), \qquad \operatorname{tr} Z \leq 2t + \delta, \tag{91}$$

and the fixed trace bound $\operatorname{tr} Z \leq 2B_0$ holds exactly by construction. When a test yields both a positive certificate and an approximate output, either conclusion is sound, and we use the certificate. In the output case, overwrite only the observed entries of Y to obtain a matrix Y'; the correction is a sum of s single-entry matrices, each of nuclear norm at most δ, so

$$\|Y' - Y\|_* \leq s\delta, \qquad \|Y'\|_* \leq t + (s + 1/2)\delta = t + \varepsilon/4. \tag{92}$$

This bound is computed from t and ε alone, so the nuclear norm of Y' never needs to be evaluated.

For the bisection, initialize the lower and upper bounds $\ell_0 = 0$ and $u_0 = B_0$, retain $Y^{(0)}$ as a feasible candidate whose norm is at most u_0, and write $w = u - \ell$ for the current interval width. At each test, set $t = (\ell + u)/2$. A positive infeasibility certificate proves $t < \mathrm{OPT}_{\mathrm{comp}}$, so replace ℓ by t; otherwise form the repaired matrix in (92) and, if $t + \varepsilon/4 < u$, retain it as the new candidate and replace u by $t + \varepsilon/4$. In all other cases retain the previous upper bound and candidate. In that case $u \leq t + \varepsilon/4 = \ell + w/2 + \varepsilon/4$, so the unchanged width already satisfies $w \leq w/2 + \varepsilon/4$. These updates preserve

$$\ell \leq \mathrm{OPT}_{\mathrm{comp}} \leq u,$$

and the retained candidate is always feasible with nuclear norm at most u. Every outcome therefore gives

$$w_{\mathrm{new}} \leq \frac{w}{2} + \frac{\varepsilon}{4},$$

and after J tests

$$w_J \leq 2^{-J} B_0 + \frac{\varepsilon}{2}(1 - 2^{-J}) \leq \varepsilon.$$

The retained candidate then satisfies (88), which establishes both the accuracy and the stated test count. □

The completion result is an optimization statement for the convex nuclear-norm problem that needs no distributional assumption on the observed entries. Identifying an underlying low-rank matrix from those entries requires additional hypotheses, and the entry repair can increase rank. Since the complete procedure consists of several fixed-loss GD runs followed by explicit matrix corrections, its exact-real operation counts establish neither finite-precision stability nor a practical speed advantage over other completion methods.

I THE EXACT LOG-DETERMINANT COMPARISON

The polynomial regularizer of Appendix G was analyzed by comparison with the exact barrier $-\mu \log\det X$, and this appendix records the analysis of that exact loss in its own right. The same lower bound on the smallest eigenvalue and the same trial-descent estimate give a direct square-factor gradient-descent guarantee, with a larger step size because no polynomial truncation error has to be absorbed. The weighted inequality of Lemma G.6 sharpens both the contraction factor and the gradient certificate by a factor two relative to Theorem G.1. The price is that the gradient of this loss contains $V^{-\top}$, so each step needs a matrix inverse or linear solves, which is why the main theorem uses the polynomial construction instead.

I.1 A GENERAL CONVEX OBJECTIVE

We keep the notation of Appendix G: $\mathbb{S}^d$ is the space of real symmetric $d \times d$ matrices, $\mathcal{D}_d$ the set of density matrices of size $d \geq 2$ defined there (positive semidefinite matrices of unit trace), and $\langle A, H\rangle = \operatorname{tr}(AH)$ the Frobenius inner product. Suppose, as before, that the objective ϕ is convex and twice continuously differentiable on an open convex neighborhood of $\mathcal{D}_d$, with a gradient bound $L > 0$ and a matrix smoothness constant $M \geq 0$ satisfying

$$\|\nabla\phi(X)\|_{\mathrm{op}} \leq L, \qquad 0 \leq \nabla^2\phi(X)[H,H] \leq M\|H\|_F^2 \quad (X \in \mathcal{D}_d,\ H \in \mathbb{S}^d). \tag{93}$$

Fix a barrier weight $\mu > 0$ and, on the invertible square factors, define

$$X(V) = \frac{VV^\top}{\|V\|_F^2}, \qquad \mathcal{L}_\mu(V) = \phi(X(V)) - \mu \log\det X(V). \tag{94}$$

Every positive definite matrix in $\mathcal{D}_d$ has such a factor, the normalization enforces the trace constraint through the readout, and the barrier acts on the whole positive semidefinite cone, so that any additional affine constraints enter through ϕ alone.

Theorem I.1 (Euclidean gradient descent on a square factor)**.** *Under* (93)*, set*

$$T = L + \mu d, \qquad r = \frac{\mu}{4T}, \qquad 0 < \eta \leq \frac{1}{32T + 4M}, \qquad h_{\min} = \frac{\eta}{1 + 4\eta L}. \tag{95}$$

Starting from the scaled-identity initialization $V_0 = I_d/\sqrt{d}$, perform ordinary Euclidean gradient descent with the constant step size η,

$$V_{t+1} = V_t - \eta\nabla\mathcal{L}_\mu(V_t). \tag{96}$$

Every iterate is invertible, and for all $t \geq 0$

$$X(V_t) \succeq rI_d, \qquad 1 \leq \|V_t\|_F^2 \leq 1 + 4\eta L. \tag{97}$$

The matrix objective $\psi(X) = \phi(X) - \mu\log\det X$ has a unique minimizer on the positive definite part of $\mathcal{D}_d$, and if its value is $\mathcal{L}_\mu^*$, then

$$\mathcal{L}_\mu(V_t) - \mathcal{L}_\mu^* \leq 2L(1 - 8h_{\min}\mu)^t. \tag{98}$$

Moreover, with $W_t = V_t/\|V_t\|_F$ and $G_t = \|V_t\|_F\nabla\mathcal{L}_\mu(V_t)$,

$$\phi(X(V_t)) - \min_{X \in \mathcal{D}_d} \phi(X) \leq \mu(d-1) + \frac{\|G_t\|_F^2}{16\mu}. \tag{99}$$

The factor stores d^2 real coordinates, and the update uses the gradient of one fixed loss throughout the trajectory.

Put differently, the exact barrier turns the trajectory from the scaled-identity start into a run of gradient descent on a region where the loss is gradient dominated, even though the loss itself is invariant under nonzero scalings and right orthogonal transformations of the factor and so is never strongly convex in the factor coordinates. Normalized factors appear below only as a device for analyzing the Euclidean update.

For an invertible V, put $s = \|V\|_F^2$ for the squared radius, $W = V/\sqrt{s}$ for the normalized factor, $X = WW^\top$, $A = \nabla\phi(X)$, and $\lambda = \langle A, X\rangle$. Direct differentiation gives

$$\nabla\mathcal{L}_\mu(V) = \frac{2}{s}(A - \lambda I_d + \mu d I_d)V - 2\mu V^{-\top}, \tag{100}$$

whose evaluation requires a matrix inverse or linear solves. On the unit Frobenius sphere define

$$G(W) = 2\big[(A - \lambda I_d + \mu d I_d)W - \mu W^{-\top}\big], \tag{101}$$

the gradient of $W \mapsto \psi(WW^\top)$ along that sphere; it is tangent to the sphere and determines the raw gradient by homogeneity,

$$\langle W, G(W)\rangle = 0, \qquad \nabla\mathcal{L}_\mu(V) = G(W)/\sqrt{s}. \tag{102}$$

The first lemma is the exact-barrier analogue of Lemma G.3: the barrier keeps every point of a trial step, not only its endpoint, above the eigenvalue level r, which is what permits a uniform descent estimate without any smoothness assumption near singular factors. The admissible step range $1/(16T)$ is four times that of the polynomial lemma, since no truncation error has to be absorbed.

Lemma I.2 (An invariant lower bound on the smallest eigenvalue)**.** *Suppose $\|W\|_F = 1$ and $WW^\top \succeq rI_d$, with T, r as in (95). For every $0 \le h \le 1/(16T)$, the normalized point*

$$W(h) = \frac{W - hG(W)}{\|W - hG(W)\|_F}$$

is well defined and satisfies $W(h)W(h)^\top \succeq rI_d$.

Proof. Because $X \succeq rI_d$ and $|\lambda| \le L$, the two terms of $G(W)$ are bounded by $2(2L + \mu d)$ and by $2\mu\|W^{-\top}\|_F = 2\mu\sqrt{\operatorname{tr} X^{-1}} \le 2\mu\sqrt{d/r}$, so

$$\|G(W)\|_F \le 2(2L + \mu d) + 2\mu\sqrt{d/r} \le 8T, \tag{103}$$

where the last inequality uses $\mu d \le T$ and $\mu\sqrt{d/r} = 2\sqrt{\mu d T} \le 2T$. Set $D = I_d - 2h(A - \lambda I_d + \mu d I_d)$ and $a = 1 - 4hT$, so that $D \succeq aI_d$ and $a \ge 3/4$ for $h \le 1/(16T)$. Since $W^{-\top} = X^{-1}W$, the unnormalized point $Y = W - hG(W)$ obeys

$$\begin{aligned} YY^\top &= DXD + 4h\mu D + 4h^2\mu^2 X^{-1} \\ &\succeq (ra^2 + 4h\mu a)I_d, \end{aligned}$$

where $DXD \succeq rD^2 \succeq ra^2 I_d$ requires no commutation of D and X. Tangency in (102) gives

$$\|Y\|_F^2 = 1 + h^2\|G(W)\|_F^2 \le 1 + 64h^2T^2,$$

and substituting $r = \mu/(4T)$,

$$ra^2 + 4h\mu a - r(1 + 64h^2T^2) = 2h\mu - 28h^2\mu T \ge 0,$$

the last inequality because $28hT \le 28/16 < 2$. Dividing the matrix lower bound by $\|Y\|_F^2$ gives $W(h)W(h)^\top \succeq rI_d$. □

The decrease estimate below plays the role of Lemma G.5; its step threshold $1/(32T + 4M)$ is twice the polynomial one because the barrier Hessian is bounded by $8T$ on the invariant region and no truncation term appears.

Lemma I.3 (Uniform decrease along normalized trial steps)**.** *Under the hypotheses of Lemma I.2, every $0 \le h \le (32T + 4M)^{-1}$ satisfies*

$$\psi(W(h)W(h)^\top) \le \psi(WW^\top) - \frac{h}{2}\|G(W)\|_F^2. \tag{104}$$

Proof. We bound the derivatives of the ambient loss on the invariant region and then differentiate along the trial path, which stays in that region by Lemma I.2. On invertible factors define the ambient function

$$F(U) = \phi(UU^\top) - 2\mu\log|\det U|,$$

which agrees with $\psi(UU^\top)$ when $\|U\|_F = 1$. At such a point with $UU^\top \succeq rI_d$, the chain rule yields

$$\nabla^2(\phi \circ \mathrm{Gram})(U)[H,H] = 2\langle \nabla\phi(UU^\top), HH^\top\rangle + \nabla^2\phi(UU^\top)[UH^\top + HU^\top, UH^\top + HU^\top],$$

whose absolute value is at most $(2L + 4M)\|H\|_F^2$, while the barrier Hessian has operator norm at most $2\mu\|U^{-1}\|_{\mathrm{op}}^2 \le 2\mu/r = 8T$. Together with the gradient,

$$\|\nabla^2 F(U)\|_{\mathrm{op}} \le 10T + 4M, \qquad \|\nabla F(U)\|_F \le 2L + 2\mu\sqrt{d/r} \le 6T,$$

and because $h \le (32T + 4M)^{-1} \le 1/(16T)$, Lemma I.2 makes these bounds valid at every point of the trial path.

Put $g = \|G(W)\|_F$. If $g > 0$, tangency gives the unit-circle parametrization

$$W(h) = \cos\theta(h) W - \sin\theta(h)\frac{G(W)}{g}, \qquad \theta(h) = \arctan(hg),$$

and the bounds $|\theta'| \le g$ and $|\theta''| \le 2g^2$ imply $\|W'(h)\|_F \le g$ and $\|W''(h)\|_F \le 3g^2$; the same conclusions hold for the constant path when $g = 0$. The second derivative of $F(W(h))$ is therefore at most $(28T + 4M)g^2$, while its first derivative at zero is $-g^2$, so Taylor's theorem gives

$$F(W(h)) \le F(W) - hg^2 + \frac{h^2}{2}(28T + 4M)g^2 \le F(W) - \frac{h}{2}g^2,$$

which is (104). □

Gradient domination and the certificate come next. Both hold at every invertible unit factor with no lower bound on the smallest eigenvalue, which is what allows Proposition I.6 below to be stated at arbitrary invertible factors rather than only along the trajectory.

Lemma I.4 (Gradient domination and an unregularized certificate)**.** *The objective ψ has a unique minimizer X_* on the positive definite part of $\mathcal{D}_d$. At every invertible unit factor with $X = WW^\top$,*

$$\|G(W)\|_F^2 \ge 16\mu\,[\psi(X) - \psi(X_*)]. \tag{105}$$

At the same point,

$$\lambda - \lambda_{\min}(A) \le \mu(d-1) + \frac{\|G(W)\|_F^2}{16\mu}. \tag{106}$$

Proof. Existence and uniqueness of X_* come from strong convexity, the gradient inequality from the weighted log-determinant inequality, and the certificate from a decomposition of the matrix gradient tested against unit vectors. The barrier diverges at the boundary of $\mathcal{D}_d$ while ϕ is continuous on this compact set, so ψ attains its minimum in the positive definite part. Since every eigenvalue of a positive definite $X \in \mathcal{D}_d$ is at most one,

$$\nabla^2(-\mu \log\det X)[H,H] = \mu\,\mathrm{tr}(X^{-1}HX^{-1}H) \ge \mu\|H\|_F^2,$$

and convexity of ϕ then makes ψ μ-strongly convex along the trace-one domain, which also proves that X_* is unique.

Remove the scalar part of the matrix gradient by setting

$$S = \nabla\psi(X) - \langle \nabla\psi(X), X\rangle I_d.$$

For every positive definite $Y \in \mathcal{D}_d$, convexity of ϕ and Lemma G.6 give

$$\begin{aligned}\psi(X) - \psi(Y) &\le \langle S, X - Y\rangle - \mu D_{-\log\det}(Y, X)\\ &\le \|SX^{1/2}\|_F\|(X-Y)X^{-1/2}\|_F - \mu\|(X-Y)X^{-1/2}\|_F^2\\ &\le \frac{\|SX^{1/2}\|_F^2}{4\mu} = \frac{\|G(W)\|_F^2}{16\mu},\end{aligned}$$

where the second line is the weighted Cauchy–Schwarz inequality and the third maximizes a scalar quadratic; taking $Y = X_*$ proves (105).

For the certificate, the same matrix S satisfies

$$A = (\lambda - \mu d) I_d + \mu X^{-1} + S, \qquad \operatorname{tr}(SX) = 0, \qquad \operatorname{tr}(SXS) = \frac{\|G(W)\|_F^2}{4}. \tag{107}$$

For a unit vector u, put $a = u^\top X^{-1} u \geq 1$. Because $\operatorname{tr}(SX) = 0$ and $\|(uu^\top - X)X^{-1/2}\|_F^2 = a - 1$ by expansion, the weighted Cauchy–Schwarz inequality gives

$$|u^\top S u| = |\langle S, uu^\top - X \rangle| \leq \frac{\|G(W)\|_F}{2} \sqrt{a - 1},$$

so that, minimizing over $a \geq 1$,

$$u^\top (\mu X^{-1} + S) u \geq \mu a - \frac{\|G(W)\|_F}{2} \sqrt{a - 1} \geq \mu - \frac{\|G(W)\|_F^2}{16\mu}.$$

Substituting into (107) shows that every eigenvalue of A is at least $\lambda - \mu(d-1) - \|G(W)\|_F^2/(16\mu)$, which is (106). Neither estimate used a lower bound on the smallest eigenvalue of X; that invariant is needed only to justify the constant step size. □

The accompanying script `check_sharper_bounds.py` exercises Lemma G.6 and the two inequalities of Lemma I.4 on one thousand random instances each, as a check on the constants rather than as evidence for the statements.

Proof of Theorem I.1. The normalized trial steps above describe the direction of ordinary Euclidean gradient descent, so what remains is to bound the changing radius and obtain a uniform effective step size. Let $s_t = \|V_t\|_F^2$, $W_t = V_t/\sqrt{s_t}$, and $h_t = \eta/s_t$, so that equation (102) gives

$$W_{t+1} = \frac{W_t - h_t G(W_t)}{\|W_t - h_t G(W_t)\|_F}, \qquad s_{t+1} = s_t + \frac{\eta^2}{s_t} \|G(W_t)\|_F^2. \tag{108}$$

Initially $s_0 = 1$ and $X_0 = I_d/d \succeq r I_d$ because $r \leq 1/(4d)$, and since the radius is nondecreasing, $h_t \leq \eta \leq 1/(16T)$, so induction with Lemma I.2 proves the eigenvalue bound and invertibility for all iterates. Applying Lemma I.3 at step t and substituting the radius identity yields

$$s_{t+1} - s_t \leq 2\eta \, [\psi(X_t) - \psi(X_{t+1})]. \tag{109}$$

The barrier is minimized over $\mathcal{D}_d$ at I_d/d, and the operator norm bound on $\nabla\phi$ gives $|\phi(X) - \phi(Y)| \leq L\|X - Y\|_* \leq 2L$ for density matrices, where $\|\cdot\|_*$ is the nuclear norm, so

$$\psi(X_0) - \psi(X_*) \leq 2L.$$

Telescoping (109) proves $s_t \leq 1 + 4\eta L$ and hence $h_t \geq h_{\min}$, after which sufficient decrease and (105) give

$$\psi(X_{t+1}) - \psi(X_*) \leq (1 - 8 h_{\min} \mu)[\psi(X_t) - \psi(X_*)],$$

which iterates to (98). Finally, convexity of ϕ bounds the objective gap by the Frank–Wolfe gap,

$$\phi(X) - \min_{Y \in \mathcal{D}_d} \phi(Y) \leq \max_{Y \in \mathcal{D}_d} \langle A, X - Y \rangle = \lambda - \lambda_{\min}(A),$$

and Lemma I.4 turns this into (99). □

I.2 Affine constraints and a gradient-based certificate

The theorem applies to the bounded affine systems of Corollary 4.2, and we repeat the slack-coordinate embedding of Appendix G in the present notation so that the constants of the two losses can be compared. Suppose the original variable is $Z \in \mathbb{S}^n$, that the constraints include $Z \succeq 0$ and a justified bound $\operatorname{tr} Z \leq R$ with $R > 0$, and that the other constraints read

$$b_i + \langle C_i, Z \rangle \leq 0, \qquad 1 \leq i \leq m,$$

with symmetric C_i, an equality being represented by its two inequality signs. The residuals may be rescaled by specified positive constants before the construction, and the accuracy guarantee then refers to those residual units.

Set $d = n + 1$ and read out $Z = RX_{1:n,1:n}$ from $X \in \mathcal{D}_d$. Adding this one slack coordinate represents all positive semidefinite matrices with trace at most R and nothing else, because for every such Z the matrix

$$X = \operatorname{diag}\big(Z/R, 1 - \operatorname{tr} Z/R\big)$$

is a density matrix. Define the embedded coefficient matrices and residuals

$$B_i = R \operatorname{diag}(C_i, 0), \qquad a_i(X) = b_i + \langle B_i, X\rangle, \qquad a_0(X) = 0, \tag{110}$$

with $B_0 = 0$ and $b_0 = 0$. The zero residual ensures that the maximum constraint violation and its minimum value,

$$v(X) = \max_{0\le i\le m} a_i(X), \qquad \gamma = \min_{X\in\mathcal{D}_d} v(X), \tag{111}$$

are nonnegative, and compactness shows that the original trace-bounded system is feasible if and only if $\gamma = 0$.

Let $q = m + 1 \ge 2$ count the residuals and take any positive bound $L \ge \max_i \|B_i\|_{\mathrm{op}}$; the main statement uses the maximum itself when it is positive, and if all B_i vanish the residuals are constant and feasibility can be checked directly. For a temperature $\tau > 0$ define the log-sum-exp smoothing and its softmax weights,

$$\phi_\tau(X) = \tau \log \sum_{i=0}^{m} \exp(a_i(X)/\tau), \qquad p_i(X) = \frac{\exp(a_i(X)/\tau)}{\sum_{j=0}^{m} \exp(a_j(X)/\tau)}. \tag{112}$$

The function is convex and smooth, its gradient is $A = \sum_i p_i B_i$, so that $\|A\|_{\mathrm{op}} \le L$, and its matrix Hessian satisfies

$$\nabla^2\phi_\tau(X)[H, H] = \frac{1}{\tau} \operatorname{Var}_{i\sim p}\langle B_i, H\rangle \le \frac{dL^2}{\tau}\|H\|_F^2, \tag{113}$$

so Theorem I.1 applies with $M = dL^2/\tau$. Substituting that constant into (95) would cost a factor d in the step size, and the lemma below recovers it by bounding the factor Hessian directly, in the same way as the proof of Theorem G.8.

Lemma I.5 (Step size for the log-sum-exp smoothing)**.** *For $\phi = \phi_\tau$, every conclusion of Theorem I.1 remains valid under the step-size bound*

$$0 < \eta \le \frac{1}{32(L + \mu d) + 4L^2/\tau}. \tag{114}$$

Proof. At a unit factor W, the Hessian of $W \mapsto \phi_\tau(WW^\top)$ is the sum of the operator $H \mapsto 2AH$ and a positive semidefinite covariance operator, namely $1/\tau$ times the covariance of the factor gradients $2B_iW$ under the weights p_i. The operator norm of the latter is bounded by its trace, which is at most

$$\frac{1}{\tau}\sum_i p_i \|2B_iW\|_F^2 \le \frac{4L^2}{\tau},$$

so the ambient factor Hessian has norm at most $2L + 4L^2/\tau$ at every point of the normalized trial path. Replacing $4M$ by $4L^2/\tau$ in the proof of Lemma I.3 gives the decrease estimate for $h \le (32T + 4L^2/\tau)^{-1}$, and since the eigenvalue bound, the gradient-domination lemma, and the radius identity do not involve M, the proof of Theorem I.1 goes through unchanged. □

The certificate that follows compares the largest constraint violation with a lower bound on its best achievable value, using only the softmax weights and the gradient norm that the loss already provides; because Lemma I.4 needs no eigenvalue bound, it is valid at every invertible factor and not only along the trajectory.

Proposition I.6 (A gradient certificate for bounded affine systems)**.** *Let V be any invertible square factor, for instance an iterate covered by Lemma I.5, and let $X = X(V)$, $W = V/\|V\|_F$, $G = G(W)$, and $A = \sum_i p_i(X)B_i$. Define*

$$\alpha = \langle A, X\rangle - \mu(d-1) - \frac{\|G\|_F^2}{16\mu}, \qquad \underline{d} = \sum_i p_i(X) b_i + \alpha. \tag{115}$$

Then $A - \alpha I_d \succeq 0$ and

$$\underline{d} \le \gamma \le v(X), \qquad v(X) - \underline{d} \le \tau \log q + \mu(d-1) + \frac{\|G\|_F^2}{16\mu}. \tag{116}$$

The norm in this formula can be computed in the raw variables as $\|G\|_F = \|V\|_F \|\nabla \mathcal{L}_\mu(V)\|_F$.

Proof. Lemma I.4 gives $A \succeq \alpha I_d$ at every invertible unit factor, so for every $Y \in \mathcal{D}_d$,

$$v(Y) \ge \sum_i p_i(X) a_i(Y) = \sum_i p_i(X) b_i + \langle A, Y \rangle \ge \underline{d},$$

and minimizing over Y proves the lower certificate. If $H(p) = -\sum_i p_i \log p_i$ denotes the entropy of the weights, the softmax identity gives

$$v(X) \le \phi_\tau(X) = \sum_i p_i(X) b_i + \langle A, X \rangle + \tau H(p), \qquad H(p) \le \log q,$$

and subtracting (115) proves the gap bound; the raw-gradient formula is (102). □

I.3 Finite-time accuracy and arithmetic scope

Choosing the smoothing and barrier parameters so that the certificate controls every affine residual gives the exact-barrier counterpart of the stopping time in Theorem G.8. Fix $0 < \varepsilon \le L$ and set

$$\mu = \frac{\varepsilon}{4d}, \qquad \tau = \frac{\varepsilon}{4 \log q}, \qquad T = L + \mu d, \qquad r = \frac{\mu}{4T}, \qquad h_{\min} = \frac{\eta}{1 + 4\eta L}, \tag{117}$$

with η satisfying (114). The stopping condition

$$\|V_t\|_F^2 \|\nabla \mathcal{L}_\mu(V_t)\|_F^2 \le 8\mu\varepsilon \tag{118}$$

makes the right side of (116) at most $\tau \log q + \mu d + \varepsilon/2 = \varepsilon$.

Finite-time accuracy for the exact log-determinant loss. Let $\Delta_t = \psi(X_t) - \psi(X_*)$. Lemma I.3 applied at W_t with the trial step $h_{\min} \le \eta$, whether or not this is the step the algorithm takes, produces a point of value at least $\psi(X_*)$, so by (98) and $1 - x \le e^{-x}$,

$$\|G(W_t)\|_F^2 \le \frac{2\Delta_t}{h_{\min}} \le \frac{4L}{h_{\min}} \exp(-8 h_{\min} \mu t).$$

The majorant on the right decreases in t and drops below $8\mu\varepsilon$ once $t \ge \log(L/(2h_{\min}\mu\varepsilon))/(8h_{\min}\mu)$, so the stopping condition holds for every $t \ge N$, where

$$N = \left\lceil \frac{1}{8 h_{\min} \mu} \max\left\{0, \log \frac{L}{2 h_{\min} \mu \varepsilon}\right\} \right\rceil; \tag{119}$$

the maximum is never active, because $h_{\min} \le 1/(32L)$ and $\mu \le \varepsilon/8$ give $L/(2h_{\min}\mu\varepsilon) \ge 128(L/\varepsilon)^2 \ge 128$. At the largest step size in (114), $T \le 5L/4$ and

$$h_{\min}^{-1} = \eta^{-1} + 4L \le 44L + \frac{16L^2 \log q}{\varepsilon},$$

so the explicit bound (119) is

$$\widetilde{O}\big(d(L/\varepsilon)^2 \log q\big), \tag{120}$$

where the suppressed factors are logarithmic in the displayed parameters and in d; a request with $\varepsilon > L$ can instead use accuracy L.

At termination, $v(X) - \underline{d} \le \varepsilon$. If $v(X) \le \varepsilon$, the readout $Z = RX_{1:n,1:n}$ is positive semidefinite, has trace at most R, and violates no chosen affine constraint by more than ε, whereas if $v(X) > \varepsilon$, then $\underline{d} > 0$ proves that the trace-bounded affine system is infeasible. This is the feasibility alternative for the exact log-determinant loss. □

The certificate also has a direct reading in the original matrix variable. Put $\overline{C} = \sum_i p_i C_i$ with $C_0 = 0$. The embedded matrix $A = R\,\mathrm{diag}(\overline{C}, 0)$ has a zero bottom diagonal entry, so $A \succeq \alpha I_d$ forces $\alpha \le 0$, and with $y = -\alpha/R \ge 0$ the certificate becomes

$$\overline{C} + yI_n \succeq 0, \qquad \underline{d} = \sum_i p_i b_i - Ry.$$

Every feasible Z would then satisfy

$$0 \ge \sum_i p_i(b_i + \langle C_i, Z\rangle) \ge \sum_i p_i b_i - y\,\mathrm{tr}\,Z \ge \underline{d},$$

which is why a positive lower certificate proves infeasibility.

These are iteration bounds in exact real arithmetic. Each step evaluates the gradient of the fixed loss, including a matrix inverse or linear solves and the softmax weights, so dense evaluation uses $O(md^2 + d^3)$ scalar arithmetic operations and q exponentials, while the factor stores d^2 entries and the input matrices need their own storage. Bounds on floating-point errors and on the cost of transcendental functions would require a separate numerical analysis. The loss is smooth on the invertible factors, but its derivatives become unbounded near singular factors, and it is the invariant-region argument above, not global smoothness, that controls the trajectory.

The trace bound must either be valid for the original problem or be an explicitly imposed restriction, since if it excludes all feasible matrices the certificate concerns only the restricted system. No strict-feasibility assumption is imposed on the original affine constraints, whose feasible set may consist entirely of singular matrices; in that case the positive definite iterates provide approximate affine feasibility rather than exact constraint repair. The theorem likewise supplies no exact constrained optimum, no polynomial bit-complexity guarantee, and no bound on the coefficients of a sum-of-squares certificate.

I.4 A LOCAL CONDITIONING LIMIT

The improved rate retains a factor of $d(L/\varepsilon)^2$, and the example below shows that this scale appears in the local conditioning of the exact log-determinant loss. Its conclusion concerns arbitrary nearby initializations at a minimizer, whereas the scaled-identity trajectory of Theorem I.1 has additional symmetry in this example.

Proposition I.7 (Two local curvature scales)**.** *Let $d \ge 4$, $0 < \varepsilon \le L$, $\mu = \varepsilon/(4d)$, and $\tau = \varepsilon/4$. There is a convex log-sum-exp smoothing of two affine functions with coefficient operator norms at most $L/2$ such that the normalized exact log-determinant loss has two positive tangent Hessian eigenvalues whose ratio is*

$$\kappa = 1 + 8a^2 d(L/\varepsilon)^2, \qquad \frac{1}{8} \le a \le \frac{1}{4}.$$

For the linearized ordinary-GD map, the best scalar step has worst contraction factor at least $(\kappa - 1)/(\kappa + 1)$ on these two modes.

Proof. The construction places the minimizer at a diagonal matrix with four large and $d - 4$ small eigenvalues, and then aligns the smoothed pair of affine functions with two of the large eigenvalues so that its covariance stiffens one tangent direction and leaves a neighboring one untouched. Set $\delta = 2\mu/L$, $a = [1 - (d-4)\delta]/4$, and

$$X_* = \mathrm{diag}(a, a, a, a, \delta, \dots, \delta), \qquad C = \mu X_*^{-1} - \frac{\mu}{a} I_d, \qquad B = \frac{L}{2}\,\mathrm{diag}(1, -1, 0, \dots, 0).$$

The assumptions give $1/8 \le a \le 1/4$, and since the first four diagonal entries of C vanish while the remaining ones lie in $[0, L/2]$, both $\|C + B\|_{\mathrm{op}}$ and $\|C - B\|_{\mathrm{op}}$ are at most $L/2$. Define

$$\phi(X) = \langle C, X\rangle + \tau \log\bigl(2\cosh(\langle B, X\rangle/\tau)\bigr).$$

At X_* the two softmax weights are equal and $\nabla\phi(X_*) = C$, so $\nabla(\phi - \mu \log\det)(X_*) = -(\mu/a) I_d$, and strict convexity makes X_* the unique minimizer over density matrices.

Let $W_* = X_*^{1/2}$. The unit tangent directions

$$H_{\text{slow}} = \frac{1}{\sqrt{2}}\,\mathrm{diag}(0,0,1,-1,0,\dots,0), \qquad H_{\text{stiff}} = \frac{1}{\sqrt{2}}\,\mathrm{diag}(1,-1,0,0,\dots,0)$$

are Hessian eigenvectors, because the constrained matrix gradient at X_* is scalar and the sphere multiplier therefore cancels the first-order factor curvature. In either direction the matrix perturbation is $D = W_* H^\top + H W_*^\top = 2\sqrt{a}\,H$, to which the log-determinant Hessian contributes $\mu\,\mathrm{tr}(X_*^{-1} D X_*^{-1} D) = 4\mu/a$, while the softmax covariance, which is rank one on the diagonal tangent space, contributes zero in the slow direction and $2aL^2/\tau$ in the stiff one. Hence

$$\lambda_{\text{slow}} = \frac{4\mu}{a}, \qquad \lambda_{\text{stiff}} = \frac{4\mu}{a} + \frac{2aL^2}{\tau}, \qquad \frac{\lambda_{\text{stiff}}}{\lambda_{\text{slow}}} = 1 + \frac{a^2 L^2}{2\mu\tau},$$

and substituting μ and τ gives the stated ratio. A scalar-step linearized update has factors $1 - h\lambda_{\text{slow}}$ and $1 - h\lambda_{\text{stiff}}$ on the two modes, and minimizing their maximum absolute value over h gives $(\kappa - 1)/(\kappa + 1)$; changing the Frobenius radius divides both Hessian eigenvalues by the same scalar and so preserves their ratio. □

The scaled-identity initialization preserves equality of the third and fourth factor coordinates in this example and therefore never excites the slow mode. Proposition I.7 is thus a local conditioning limit for the exact loss rather than a matching iteration lower bound for the scaled-identity trajectory or for the polynomial loss of Corollary 4.2.

I.5 Applications with an explicit trace bound

Since the theorem is stated for residuals, we close by recording how an optimization problem over a bounded semidefinite relaxation is written in that form, and by working one continuous example through with explicit constants. For a linear objective $h(Z) = \langle C, Z\rangle$, a threshold c is represented by the residual $a_{\text{obj}}(Z) = (h(Z) - c)/s_{\text{obj}}$ with a specified positive scale s_{obj}, so that a successful approximate output has

$$h(Z) \le c + s_{\text{obj}}\varepsilon$$

along with the reported constraint residual bounds. A positive lower certificate proves the exact threshold system infeasible and hence lower-bounds the exact constrained optimum; because an approximate output can violate the other affine constraints, its objective value alone need not upper-bound that optimum.

For the continuous example, consider the Peng–Wei relaxation from (150) with n data points and $1 \le K \le n$ clusters, whose constraints

$$Z \succeq 0, \qquad Z\mathbf{1} = \mathbf{1}, \qquad \mathrm{tr}\, Z = K, \qquad Z_{ij} \ge 0$$

supply the known trace bound $R = K$. Adding the slack coordinate gives dimension $d = n + 1$ and a square factor with $(n+1)^2$ entries, so this application treats the relaxation of Appendix L.9 through a direct trace representation.

To state one explicit residual normalization, take squared distances $0 \le d_{ij} \le 1$ and put $h(Z) = (2n)^{-1}\sum_{i,j} d_{ij} Z_{ij}$. For an objective threshold c, use both signs of each equality residual and the inequalities

$$\frac{(Z\mathbf{1})_i - 1}{K\sqrt{n}}, \qquad \frac{\mathrm{tr}\, Z - K}{K}, \qquad -\frac{Z_{ij}}{K}\ (i < j), \qquad \frac{h(Z) - c}{K}, \tag{121}$$

where the displayed equality residuals stand for their positive and negative versions and positive semidefiniteness already enforces nonnegativity of the diagonal. After the embedding, every affine coefficient in (121) has operator norm at most one: the row-equation coefficient $(e_i\mathbf{1}^\top + \mathbf{1}e_i^\top)/(2\sqrt{n})$ has norm at most one, the trace coefficient has norm one, an off-diagonal coefficient has norm $1/2$, and the objective coefficient $(d_{ij})/(2n)$ has norm at most $1/2$. One may therefore take $L = 1$ and $q = 2n + 4 + \binom{n}{2}$, the zero residual included.

For $0 < \varepsilon \le 1$, the resulting Euclidean GD trajectory reaches its gradient stopping condition within $\widetilde{O}((n+1)\varepsilon^{-2}\log q)$ steps, and an approximate output obeys

$$|(Z\mathbf{1})_i - 1| \le K\sqrt{n}\,\varepsilon, \quad |\,\mathrm{tr}\, Z - K| \le K\varepsilon, \quad Z_{ij} \ge -K\varepsilon, \quad h(Z) \le c + K\varepsilon,$$

while being positive semidefinite with trace at most K by construction; alternatively, a positive lower certificate proves that the exact Peng–Wei threshold system is infeasible. This is a finite gradient-descent guarantee for a semidefinite relaxation in explicit residual units, although producing a feasible clustering from the output still requires a separate rounding argument.

LANDSCAPE BOUNDS AND APPLICATIONS

J CERTIFICATES AND NEGATIVE CURVATURE AT FIXED SEARCH RANK

This appendix proves Theorem 4.3(a) and Corollary 4.4, then studies the log-sum-exp loss without the polynomial regularizer of Appendix G, on factors whose search rank k may be smaller than the matrix dimension d. All factor Hessians use the metric induced by the Frobenius inner product. Write

$$\mathcal{D}_d = \{X \in \mathbb{S}^d \mid X \succeq 0,\ \operatorname{tr} X = 1\}, \qquad \mathcal{S}_{d,k} = \{W \in \mathbb{R}^{d\times k} \mid \|W\|_F = 1\}$$

for the set of density matrices, that is, positive semidefinite matrices of unit trace, and for the unit sphere of factors. The tangent space of the sphere at W consists of the matrices U with $\langle W, U\rangle = 0$. For a trace bound $R > 0$, take $d = n + 1$ and read out $Z = R(WW^\top)_{1:n,1:n}$, which is positive semidefinite with trace at most R. Conversely, adding one slack coordinate embeds any $Z \succeq 0$ with $\operatorname{tr} Z \le R$ as the density matrix $\operatorname{diag}(Z/R, 1 - \operatorname{tr} Z/R)$, so the unrestricted Gram domain represents the entire trace ball. At search rank k, the output has rank at most k and the factor stores dk coordinates.

J.1 A HESSIAN CERTIFICATE FOR THE CONVEX DUALITY GAP

The first step is a certificate that needs no first-order condition: a lower bound on the factor Hessian alone controls the duality gap of a convex matrix objective. Let f be convex and C^2 on a neighborhood of $\mathcal{D}_d$, and for the factor objective $g(W) = f(WW^\top)$ put

$$X = WW^\top, \qquad A = \nabla f(X), \qquad \lambda = \langle A, X\rangle, \qquad \Delta(X) = \lambda - \lambda_{\min}(A).$$

The quantity $\Delta(X)$ is the Frank–Wolfe gap of f over $\mathcal{D}_d$ (Jaggi, 2013), and convexity gives

$$0 \le f(X) - \min_{S\in\mathcal{D}_d} f(S) \le \Delta(X). \tag{122}$$

Theorem J.1 (Trace certificate at fixed search rank; Theorem 4.3(a) restated)**.** *Let $k \ge 2$ and $\eta \ge 0$, and suppose that the sphere Hessian satisfies* $\operatorname{Hess} g(W) \succeq -\eta I$ *at a factor $W \in \mathcal{S}_{d,k}$. If, at the current matrix $X = WW^\top$,*

$$\nabla^2 f(X)[H, H] \le M\|H\|_F^2 \quad \textit{for every symmetric } H,$$

then

$$\Delta(X) \le \frac{M}{k-1} + \frac{\eta}{2}, \tag{123}$$

with no first-order condition required. If also $\operatorname{grad} g(W) = 0$*, the denominator $k-1$ can be replaced by k.*

Proof. We test the Hessian lower bound on rank-one tangent directions $uz^\top$ built from a bottom eigenvector u of A, and average over z so that the curvature of f is charged only $1/(k-1)$ of its bound. The sphere gradient and Hessian satisfy

$$\operatorname{grad} g(W) = 2(A - \lambda I)W, \tag{124}$$

$$\operatorname{Hess} g(W)[U, U] = 2\langle A - \lambda I, UU^\top\rangle + \nabla^2 f(X)[WU^\top + UW^\top, WU^\top + UW^\top]; \tag{125}$$

for the second identity, the sphere geodesic with initial velocity U has acceleration $-\|U\|_F^2 W$ at W, and this term produces the shift $-\lambda I$.

Choose a unit bottom eigenvector u of A and a unit vector $z \perp W^\top u$. Then $U = uz^\top$ is a unit tangent vector, since $\langle W, uz^\top\rangle = u^\top Wz = 0$, and the same identity gives $u \perp Wz$, so that

$$H = WU^\top + UW^\top = (Wz)u^\top + u(Wz)^\top, \qquad \|H\|_F^2 = 2\|Wz\|^2.$$

Equation (125) becomes

$$\operatorname{Hess} g(W)[U, U] = -2\Delta(X) + \nabla^2 f(X)[H, H].$$

The space $(W^\top u)^\perp \subseteq \mathbb{R}^k$ has dimension $k-1$ if $W^\top u \ne 0$ and k if $W^\top u = 0$, so averaging $\|Wz\|^2$ over an orthonormal basis $\{z_j\}$ of this space gives at most $\|W\|_F^2/(k-1) = 1/(k-1)$ in both

cases, because $\sum_j \|Wz_j\|^2 \le \|W\|_F^2$. Some unit z in the space therefore satisfies $\nabla^2 f(X)[H,H] \le 2M/(k-1)$, and the Hessian lower bound at this U reads $-\eta \le -2\Delta + 2M/(k-1)$, which rearranges to (123).

Under stationarity, $(A-\lambda I)W = 0$, and if $\Delta > 0$ then u is an eigenvector of $A - \lambda I$ with the nonzero eigenvalue $-\Delta$, so multiplying by $u^\top$ gives $u^\top W = 0$ and all k right directions are available for the average. The case $\Delta = 0$ needs no argument, since the left side of (123) then vanishes. Finally, convexity gives $f(X) - f(S) \le \langle A, X - S\rangle \le \Delta(X)$ for every $S \in \mathcal{D}_d$, which proves (122). □

In words, at any point where the factor Hessian is nearly positive semidefinite, the objective is within $M/(k-1) + \eta/2$ of the convex optimum, whatever the gradient does. The constant M/k in the stationary case is attained for the Frank–Wolfe gap, so its bound is sharp there.

Proposition J.2 (Sharp stationary constant)**.** *For $d > k$ and $M > 0$, let*

$$f(X) = \frac{M}{2}\|X - I_d/d\|_F^2, \qquad W = \frac{1}{\sqrt{k}}\begin{bmatrix} I_k \\ 0 \end{bmatrix}.$$

The factor W is a second-order stationary point of the factor problem, and

$$\Delta(WW^\top) = \frac{M}{k}, \qquad f(WW^\top) - \min_{X \in \mathcal{D}_d} f(X) = \frac{M}{2}\left(\frac{1}{k} - \frac{1}{d}\right).$$

Proof. We show that W is a global minimizer of the factor problem, compute Δ from the spectrum of A, and read the Hessian off (125). By Cauchy–Schwarz on its nonzero eigenvalues, a density matrix of rank at most k has squared Frobenius norm at least $1/k$, and the displayed factor attains this bound. It therefore globally minimizes the factor objective, and the value gap is $\frac{M}{2}(1/k - 1/d)$. Direct differentiation gives $A = M(WW^\top - I_d/d)$ and $\lambda = M(1/k - 1/d)$, and since the smallest eigenvalue of A is $-M/d$, attained on the last $d-k$ coordinates, the Frank–Wolfe gap is $\Delta = M/k$ as claimed. For a tangent perturbation $U = [U_1^\top\ U_2^\top]^\top$ with $U_1 \in \mathbb{R}^{k\times k}$, equation (125) gives

$$\operatorname{Hess} g(W)[U,U] = \frac{M}{k}\|U_1 + U_1^\top\|_F^2 \ge 0,$$

while the gradient vanishes by (124) because $(A - \lambda I)W = M(WW^\top W - W/k) = 0$; hence W is second-order stationary. □

J.2 Affine feasibility and an explicit dual witness

We now specialize the certificate to the log-sum-exp smoothing of an affine feasibility system, where the dual witness becomes explicit. Let $m \ge 1$ and $q = m + 1$, consider signed affine residuals $a_i(X) = b_i + \langle B_i, X\rangle$ for $1 \le i \le m$ together with the zero residual $a_0 = 0$ (so $b_0 = 0$ and $B_0 = 0$), and suppose that the coefficient matrices satisfy $\|B_i\|_{\mathrm{op}} \le L$ with $L > 0$. For the trace-bounded formulation above one may take $B_i = R\,\mathrm{diag}(C_i, 0)$ when the original residual is $b_i + \langle C_i, Z\rangle$. With a smoothing temperature $\tau > 0$, define the maximum constraint violation, its minimum value, the smoothed objective, and the softmax weights by

$$\begin{aligned} v(X) &= \max_{0\le i\le m} a_i(X), & \gamma &= \min_{X\in\mathcal{D}_d} v(X), \\ \Phi_\tau(X) &= \tau \log \sum_{i=0}^{m} e^{a_i(X)/\tau}, & p_i &= \frac{e^{a_i(X)/\tau}}{\sum_{j=0}^{m} e^{a_j(X)/\tau}}. \end{aligned} \tag{126}$$

In this subsection $g(W) = \Phi_\tau(WW^\top)$, its matrix gradient is $A = \sum_i p_i B_i$, and the dual value of a probability vector p is

$$\mathsf{d}(p) = \sum_{i=0}^{m} p_i b_i + \lambda_{\min}(A).$$

For every probability vector p, minimizing the weighted residual $\sum_i p_i a_i$ over $\mathcal{D}_d$ shows that $\mathsf{d}(p) \le \gamma$, and compact convex minimax gives $\gamma = \max_p \mathsf{d}(p)$ without any strict feasibility assumption. Applied to Φ_τ, whose curvature is controlled by the covariance of the coefficient matrices, the trace certificate yields a computable gap between the primal and dual values.

Corollary J.3 (Softmax certificate). *If $k \geq 2$ and the sphere Hessian satisfies* $\operatorname{Hess} g(W) \succeq -\eta I$ *at W, then*

$$\mathsf{d}(p) \leq \gamma \leq v(X), \qquad v(X) - \mathsf{d}(p) \leq \tau \log q + \frac{2L^2}{\tau(k-1)} + \frac{\eta}{2}, \tag{127}$$

and at exact first-order stationarity $k-1$ may be replaced by k.

Proof. The argument of Theorem J.1 applies with the curvature bound M replaced by a direct computation of the second derivative of Φ_τ along the test directions. An entropy bound then converts the Frank–Wolfe gap into the primal–dual gap. With the unit bottom eigenvector u of A and the tangent direction $uz^\top$ from that proof, the nonlinear curvature term is

$$\nabla^2 \Phi_\tau(X)[H,H] = \frac{4}{\tau} z^\top Q z, \qquad Q = \sum_i p_i W^\top (B_i - A) u u^\top (B_i - A) W,$$

where Q is positive semidefinite because it is a weighted sum of squares. Its trace satisfies

$$\begin{aligned} \operatorname{tr} Q &= \sum_i p_i \|W^\top B_i u\|^2 - \|W^\top A u\|^2 \\ &\leq \sum_i p_i \|B_i\|_{\mathrm{op}}^2 \|W\|_F^2 \leq L^2. \end{aligned}$$

Averaging $z^\top Q z$ over an orthonormal basis of $(W^\top u)^\perp$, a space of dimension $k-1$ or k, therefore gives at most $L^2/(k-1)$, and the Hessian lower bound at the best z yields $\Delta \leq 2L^2/(\tau(k-1)) + \eta/2$. The stationary improvement to k follows as in the preceding proof. The softmax identity

$$\Phi_\tau(X) = \sum_i p_i a_i(X) + \tau H(p), \qquad H(p) = -\sum_i p_i \log p_i \leq \log q,$$

together with $v(X) \leq \Phi_\tau(X)$ and $\mathsf{d}(p) = \sum_i p_i b_i + \lambda_{\min}(A)$, gives $v(X) - \mathsf{d}(p) \leq \Delta + \tau \log q$, and combining this with the bound on Δ proves (127). □

J.3 Many negative-curvature directions

The certificate above uses a single negative-curvature direction. Corollary 4.4 strengthens this conclusion by counting such directions. Their number is governed by the trace of the positive curvature that the smoothing contributes. Write that trace, the total positive curvature available at W, as

$$\mathcal{B}(W) = \frac{4}{\tau} \left(\sum_i p_i \|B_i W\|_F^2 - \|AW\|_F^2 \right) \leq \frac{4L^2}{\tau}, \tag{128}$$

and for $a > 0$ let $r(a)$ count the eigenvalues of A that are at most $\lambda - a$, while $\nu_{-\eta}$ counts the eigenvalues of the factor Hessian strictly below $-\eta$.

Theorem J.4 (Spectral multiplicity of negative-curvature directions). *For $0 \leq \eta < 2a$, every factor point $W \in \mathcal{S}_{d,k}$ satisfies*

$$\nu_{-\eta} \geq \max \left\{ 0,\ r(a)k - 1 - \left\lfloor \frac{\mathcal{B}(W)}{2a - \eta} \right\rfloor \right\}, \tag{129}$$

and at an exact first-order stationary point $r(a)k - 1$ may be replaced by $r(a)k$.

Proof. The idea is to exhibit a large subspace on which the linear term of the Hessian is at most $-2a$, and to show that the positive curvature of the smoothing, whose trace is bounded by $\mathcal{B}(W)$, can raise only a bounded number of its directions above $-\eta$. On the tangent space, the Hessian quadratic form is

$$\operatorname{Hess} g(W)[U,U] = 2\langle A - \lambda I, UU^\top \rangle + \frac{4}{\tau} \operatorname{Var}_{i \sim p} \langle B_i W, U \rangle, \tag{130}$$

where the variance term defines a positive semidefinite operator whose trace on the ambient factor space equals $\mathcal{B}(W)$. Orthogonal compression to any subspace cannot increase that trace.

Let E be the spectral subspace of A for eigenvalues at most $\lambda - a$. The matrices whose columns lie in E form a space of dimension $r(a)k$, and intersecting it with the tangent space, which is

defined by the single linear condition $\langle W, U\rangle = 0$, loses at most one dimension. On this intersection, $\langle A - \lambda I, UU^\top\rangle \le -a\|U\|_F^2$ because each column of U lies in E, so the first term in (130) is at most $-2a\|U\|_F^2$.

A positive semidefinite operator with trace at most $\mathcal{B}(W)$ has at most $\lfloor \mathcal{B}(W)/(2a-\eta)\rfloor$ eigenvalues that are at least $2a-\eta$. The compressed variance operator is therefore strictly below $2a-\eta$ on a subspace of the intersection whose codimension is at most that number, and on that subspace the complete Hessian is strictly below $-\eta\|U\|_F^2$. The min-max principle turns this subspace into the lower bound (129). Under stationarity, $(A-\lambda I)W = 0$ forces $P_E W = 0$, since $A - \lambda I$ is invertible on E with eigenvalues at most $-a < 0$; the entire $r(a)k$-dimensional space is then tangent, which gives the improved count. □

Remark J.5 (General convex objectives). Under the assumptions of Theorem J.1, the same argument on the space $\{uz^\top \mid z \perp W^\top u\}$ gives

$$\nu_{-\eta} \ge \max\left\{0,\ k - 1 - \left\lfloor \frac{2M}{2\Delta - \eta}\right\rfloor\right\}, \qquad 0 \le \eta < 2\Delta.$$

The nonlinear curvature operator restricted to this space is positive semidefinite by convexity, and its trace is at most $2M$ because $\|H\|_F^2 = 2\|Wz\|^2$ and $\sum_j \|Wz_j\|^2 \le 1$ over an orthonormal basis $\{z_j\}$. Stationarity again permits k in place of $k-1$.

Choosing the temperature and the search rank as in Corollary 4.4 turns the count into the alternative stated in the main text.

Corollary J.6 (Negative curvature on half the search rank). *Let $\varepsilon > 0$ and choose*

$$\tau = \frac{\varepsilon}{4\log q}, \qquad k - 1 \ge \frac{32L^2 \log q}{\varepsilon^2}. \tag{131}$$

At every factor point, either $v(X) - \mathsf{d}(p) \le \varepsilon$, or at least $\lceil (k-1)/2\rceil$ eigenvalues of its Hessian are strictly below $-\varepsilon/2$.

Proof. Suppose $v(X) - \mathsf{d}(p) > \varepsilon$. Since $\tau \log q = \varepsilon/4$, the entropy inequality $v(X) - \mathsf{d}(p) \le \Delta + \tau \log q$ from the proof of Corollary J.3 gives $\Delta > 3\varepsilon/4$. Apply Theorem J.4 with $a = \Delta$, for which $r(a) \ge 1$ because the bottom eigenvalue of A equals $\lambda - \Delta$, and with $\eta = \varepsilon/2$, so that $2a - \eta > \varepsilon$ and, by (128) and (131),

$$\frac{\mathcal{B}(W)}{2a-\eta} < \frac{16L^2\log q}{\varepsilon^2} \le \frac{k-1}{2}.$$

The count in (129) is then at least $k - 1 - \lfloor (k-1)/2\rfloor = \lceil (k-1)/2\rceil$, which proves the claim. □

Proof of Corollary 4.4. The main theorem is the preceding corollary with the alternative phrased through the Frank–Wolfe gap. The counting argument above uses only $\Delta > 3\varepsilon/4$, so it proves the negative-eigenvalue alternative whenever $\Delta > 3\varepsilon/4$. If instead $\Delta \le 3\varepsilon/4$, weak duality $\mathsf{d}(p) \le \gamma$ and the softmax entropy bound give

$$v(X) - \gamma_* \le v(X) - \mathsf{d}(p) \le \tau \log q + \Delta \le \varepsilon,$$

where γ_* in the main text denotes the same optimal maximum constraint violation as γ in this appendix; this is the first alternative of the theorem. □

J.4 Isotropic perturbations and local gradient descent

Having counted negative-curvature directions, we ask what they buy an algorithm. The lemma and theorem below quantify how a large unstable subspace shortens a gradient-only escape phase started from a random perturbation, without any eigenvector being computed. The first ingredient is that a uniform perturbation places a fixed fraction of its mass on any subspace of proportional dimension, with failure probability exponentially small in that dimension.

Lemma J.7 (Perturbation alignment). *Let E_- be an s-dimensional subspace of $\mathbb{R}^D$ with $1 \le s \le D$, and let ξ be uniform on the sphere of radius $r > 0$. Then*

$$\Pr\left\{\|P_{E_-}\xi\| \ge \frac{r}{2}\sqrt{\frac{s}{D}}\right\} \ge 1 - e^{-s/16} - e^{-D/8}. \tag{132}$$

Proof. Both tails are chi-squared tails obtained from the moment-generating function. Write $\xi = rG/\|G\|$ for a standard Gaussian vector G, so that $\|P_{E_-}G\|^2$ and $\|G\|^2$ are chi-squared variables with s and D degrees of freedom. Recall that a chi-squared variable with j degrees of freedom has moment-generating function $(1-2t)^{-j/2}$ for $t < 1/2$. Markov's inequality at $t = -1/2$ gives $\Pr\{\|P_{E_-}G\|^2 < s/2\} \le e^{-s/16}$, and at $t = 1/4$ it gives $\Pr\{\|G\|^2 > 2D\} \le e^{-D/8}$. On the intersection of the two complementary events, $\|P_{E_-}\xi\|^2/r^2 = \|P_{E_-}G\|^2/\|G\|^2 \ge (s/2)/(2D) = s/(4D)$, which is (132). □

The second ingredient is a local analysis of gradient descent from such a perturbation, in which the exit time depends on the dimension only through the alignment fraction.

Theorem J.8 (Local gradient-descent escape)**.** *Let F be C^2 on a neighborhood of the closed ball of radius b in $\mathbb{R}^D$, with an ℓ-Lipschitz gradient and a ρ-Lipschitz Hessian there, where $\ell > 0$ and $\rho \ge 0$. Assume that $\nabla^2 F(0)$ has at least $s \ge 1$ eigenvalues at most $-\gamma$ for some $\gamma > 0$, put $\chi = \sqrt{s/D}$ for the alignment fraction, and suppose*

$$b \le \frac{\gamma\chi}{16\rho}, \qquad \|\nabla F(0)\| \le \frac{\gamma b \chi}{32}, \tag{133}$$

where the first restriction is omitted when $\rho = 0$. Draw x_0 uniformly from the sphere of radius $b/4$ and apply

$$x_{t+1} = x_t - \alpha\nabla F(x_t), \qquad 0 < \alpha \le 1/\ell,$$

until the iterate leaves the open ball of radius b. With probability at least $1 - e^{-s/16} - e^{-D/8}$, the exit occurs after at most

$$1 + \left\lceil \frac{\log(16/\chi)}{\log(1+\alpha\gamma)} \right\rceil \tag{134}$$

gradient steps.

Proof. Linearizing the iteration at the origin, we track the component in the unstable subspace, which grows geometrically, and bound the accumulated linearization error by a geometric series that condition (133) keeps below half of the initial alignment supplied by Lemma J.7. Let P be the orthogonal projection onto an s-dimensional invariant subspace of $\nabla^2 F(0)$ spanned by eigenvectors with eigenvalues at most $-\gamma$. Put $J = I - \alpha\nabla^2 F(0)$ and denote by J_- its restriction to this subspace, all of whose eigenvalues are at least $1 + \alpha\gamma$. While the previous iterates remain in the ball, Taylor's theorem around the origin gives

$$x_{t+1} = Jx_t + e_t, \qquad \|e_t\| \le \alpha\left(\|\nabla F(0)\| + \frac{\rho}{2}b^2\right),$$

where the error collects the gradient at the origin and the quadratic remainder. Backward iteration on the invariant subspace yields

$$J_-^{-t}Px_t = Px_0 + \sum_{j=0}^{t-1} J_-^{-j-1}Pe_j,$$

and since $\|J_-^{-j-1}\| \le (1+\alpha\gamma)^{-j-1}$, the geometric series and (133) bound the norm of the sum by

$$\frac{\|\nabla F(0)\|}{\gamma} + \frac{\rho b^2}{2\gamma} \le \frac{b\chi}{16}.$$

Lemma J.7, applied with radius $b/4$, gives $\|Px_0\| \ge b\chi/8$ with the asserted probability, and on this event the triangle inequality leaves $\|J_-^{-t}Px_t\| \ge b\chi/16$, so that

$$\|Px_t\| \ge (1+\alpha\gamma)^t \frac{b\chi}{16}.$$

Because $\|Px_t\| \le \|x_t\| < b$ inside the open ball, remaining there until the time in (134) would contradict this growth, so the iterate has exited by that time. □

The exit time therefore depends on the ambient dimension only through the alignment fraction χ, so that a larger unstable subspace shortens the escape. The step-size restriction $\alpha \le 1/\ell$ is the usual one for the surrounding descent phases and is not used in the exit argument. To apply the theorem on the factor sphere, use the normalized tangent retraction

$$\mathcal{R}_W(x) = \frac{W + x}{\sqrt{1 + \|x\|_F^2}}, \qquad \langle W, x\rangle = 0, \qquad F_W(x) = g(\mathcal{R}_W(x)),$$

which is a second-order retraction, so that the pullback gradient and Hessian of F_W at zero agree with the intrinsic derivatives of g. Ordinary GD on F_W is then a gradient-only tangent-space phase, and the lemma below supplies the Lipschitz constants that the theorem requires.

Lemma J.9 (Explicit pullback regularity)**.** *For the log-sum-exp smoothing* (126)*, valid gradient and Hessian Lipschitz constants for every pullback F_W are*

$$\ell_{\mathrm{pb}} = 14L + \frac{4L^2}{\tau}, \qquad \rho_{\mathrm{pb}} = 108L + \frac{84L^2}{\tau} + \frac{16L^3}{\tau^2}. \tag{135}$$

Proof. We bound the first three derivatives of g on the unit sphere of factors and of the normalization map on inputs of norm at least one, and then compose the bounds with the chain rule. Write $g_i(W) = a_i(WW^\top)$ for the branch objectives. Their gradients $Dg_i[h] = 2\langle B_i W, h\rangle$ have norm at most $2L$ at a unit factor, their Hessians $h \mapsto 2B_i h$ have operator norm at most $2L$ on $\mathbb{R}^{d\times k}$ with the Frobenius norm, and their third derivatives vanish. At a unit ambient factor, the first three Euclidean derivative norms of $g = \Phi_\tau \circ \mathrm{Gram}$ are therefore bounded by

$$G = 2L, \qquad H = 2L + 4L^2/\tau, \qquad T = 12L^2/\tau + 16L^3/\tau^2,$$

for the following reasons. The gradient of g is the softmax average of the branch gradients. Its Hessian adds to the averaged branch Hessians the covariance of the branch gradients divided by τ, which is bounded by $(2L)^2/\tau = 4L^2/\tau$. Its third derivative consists of three gradient–Hessian covariance terms, each bounded by $4L^2/\tau$ through Cauchy–Schwarz, and a centered third moment of the branch gradients divided by τ^2. Since $|Dg_i[h]| \le 2L$ for unit h, the centered values are at most $4L$ in absolute value and their variance is at most $4L^2$, so the third centered moment is at most $4L \cdot 4L^2 = 16L^3$, which gives the term $16L^3/\tau^2$. For the normalization map $y \mapsto y/\|y\|_F$, the derivative norms through order three are at most 1, 6, and 36 on inputs of norm at least one, and $\|W+x\|_F \ge 1$ throughout the tangent space because $x \perp W$. The chain rule for the composition then gives $H + 6G$ for the pullback Hessian bound and $T + 18H + 36G$ for the pullback third-derivative bound, which are the constants in (135). □

On the factor sphere, $D = dk - 1$, and Corollary J.6 provides s proportional to k. Theorem J.8 then replaces the local dependence on D by a dependence on D/s, which is of order d, while the alignment failure probability becomes exponentially small in k. This quantifies the advantage that many negative-curvature directions give a gradient-only escape, although it does not remove the matrix-dimension dependence of isotropic noise. The theorem guarantees only exit from a neighborhood, and because a perturbation can increase the objective before the gradient steps begin, a global iteration bound would also need a decrease test and an account of successive perturbations, as in general perturbed-GD analyses (Jin et al., 2017). The local result alone gives no uniform hitting-time theorem for unmodified gradient descent.

K Proofs for overparameterized Gram objectives

This appendix proves the unit-diagonal bound in Theorem 4.3(b), whose constants use the unit-diagonal structure directly. It then examines square search rank and the limit of normalized descent. Throughout, matrix derivatives are taken with respect to the Frobenius inner product on symmetric matrices, a factor $V \in \mathbb{R}^{N\times k}$ has unit rows $v_i \in \mathbb{R}^k$, and the product of spheres on which such factors live carries the ambient Frobenius metric.

K.1 A representation-specific landscape bound

A unit-row factor satisfies the constraints $X_{ii} = 1$ identically, so its Gram matrix lies in the *elliptope* $\mathcal{E}_N = \{X \succeq 0 \mid X_{ii} = 1\}$ of unit-diagonal positive semidefinite matrices. The factors themselves

form the product of spheres $\mathcal{M}_{N,k} = (\mathbb{S}^{k-1})^N$, where the search rank $k \geq 2$ is the number of columns. For a C^2 concave function f defined on a neighborhood of $\mathcal{E}_N$, we write $F(V) = f(VV^\top)$ for the factor objective and measure a symmetric matrix A by its linear range over the elliptope,

$$R(A) = \max_{Z \in \mathcal{E}_N} \langle A, Z \rangle - \min_{Z \in \mathcal{E}_N} \langle A, Z \rangle.$$

For linear f, the bound in Theorem 4.3(b) is exactly Mei et al. (2017, Theorem 2), including its approximate-Hessian term and absence of a first-order assumption. We extend their projected-Gaussian argument (Mei et al., 2017, Section 3.2) by retaining the nonlinear curvature term $C/2$. The theorem keeps the unit-diagonal structure throughout. Its constants $R(A)$ and C are computed from the elliptope and from the sphere tangent spaces, and for structured objectives they can be much smaller than the constants the slack-augmented trace-one embedding would produce for the same problem.

Fix $V \in \mathcal{M}_{N,k}$ and a comparator $X_* = YY^\top \in \mathcal{E}_N$, with $Y \in \mathbb{R}^{N \times s}$ having rows $y_i^\top$. Draw $G \in \mathbb{R}^{k \times s}$ with independent $\mathcal{N}(0, 1/k)$ entries, set $u_i = (I - v_i v_i^\top) G y_i$, let U have rows $u_i^\top$, and put $H = VU^\top + UV^\top$. The curvature constant $C \geq 0$ in Theorem 4.3(b) satisfies (9), with expectation over G. The constant may depend on V and X_*, while the applications establish bounds uniform over these choices.

In words, a factor whose Hessian has no tangent eigenvalue above η is within $O(1/(k-1))$ of the elliptope optimum. The constant records how far the linearization can vary over the elliptope and how much nonlinear curvature the probe sees, and the only other loss is the $N\eta/2$ paid for an inexact second-order condition.

K.2 Gaussian tangent moments and the landscape bound

The proof compares V with the comparator through a random tangent direction, so we first record the second moments of that direction. Factor the comparison matrix as $X_* = YY^\top$ with unit rows y_i, let $G \in \mathbb{R}^{k \times s}$ have independent $\mathcal{N}(0, 1/k)$ entries, where s is the number of columns of Y, and write $X = VV^\top$. With the row projections $P_i = I - v_i v_i^\top$, the tangent rows are $u_i = P_i G y_i$, the matrix U again collects them, and $H = VU^\top + UV^\top$ is the induced perturbation of the Gram matrix. Since $\mathbb{E}[(Gy_i)(Gy_j)^\top] = (X_*)_{ij} I/k$,

$$\mathbb{E}(UU^\top)_{ij} = (X_*)_{ij} \left(1 - \frac{2}{k} + \frac{X_{ij}^2}{k}\right), \qquad \mathbb{E}\|U\|_F^2 = N(1 - 1/k), \tag{136}$$

$$\mathbb{E}H_{ij}^2 = \frac{2}{k}(1 - X_{ij}^2)(1 - X_{ij}(X_*)_{ij}) \leq \frac{4}{k}. \tag{137}$$

The first line follows from $\operatorname{tr}(P_i P_j) = k - 2 + X_{ij}^2$, and for the second we put $a = X_{ij}$ and $b = (X_*)_{ij}$ and note that the two summands $v_i^\top P_j G y_j$ and $v_j^\top P_i G y_i$ of H_{ij} each have variance $(1 - a^2)/k$ and covariance $-ab(1 - a^2)/k$.

Proof of Theorem 4.3(b). We compute the Hessian of F along a tangent direction, average it over the Gaussian probe using the moments just recorded, and convert the resulting linear inequality into a value bound by concavity. The Hessian hypothesis enters when the averaged quadratic form is bounded by $\eta\, \mathbb{E}\|U\|_F^2$. Write $X = VV^\top$ and $A = \nabla f(X)$, and let Λ be the diagonal matrix with entries $\Lambda_{ii} = (AV)_i \cdot v_i$, so that $\operatorname{tr}\Lambda = \langle A, X \rangle$. Because a sphere geodesic with initial velocity U has row acceleration $-\|u_i\|^2 v_i$, differentiating F twice along it gives

$$\operatorname{Hess} F(V)[U, U] = 2\langle A - \Lambda, UU^\top \rangle + \nabla^2 f(X)[H, H]. \tag{138}$$

Taking expectations over the probe with the moments (136), and noting that the Schur product theorem places $Z = X_* \circ X \circ X$ in the elliptope, we bound the left side by $\eta N(1 - 1/k)$ and multiply by $k/2$ to obtain

$$(k-2)\langle A, X_* \rangle + \langle A, Z \rangle - (k-1)\langle A, X \rangle \leq C/2 + N\eta(k-1)/2,$$

which rearranges to

$$(k-1)\langle A, X_* - X \rangle \leq \langle A, X_* - Z \rangle + C/2 + N\eta(k-1)/2.$$

The first term on the right compares two elliptope points and is at most $R(A)$, while concavity of f gives $f(X_*) - f(X) \leq \langle A, X_* - X \rangle$, so dividing by $k - 1$ yields the bound (10). □

Uniform constants always exist for a fixed N and a fixed C^2 function on a neighborhood of the compact elliptope: if $\|\nabla^2 f(X)\|_{\mathrm{op}} \le M$, then summing the entry bound (137) over all N^2 entries gives $C \le 4MN^2$. That this generic bound grows like N^2 is the reason a dimension-independent rate needs more structure than smoothness alone, and the applications below obtain their constants from the coefficient structure of each objective instead.

K.3 Square search rank and the limits of normalized descent

The landscape bound charges $O(1/(k-1))$ for a search rank below the number of rows, and the first question is whether the charge disappears once the factor is square. It does: at $k \ge N$ an exact second-order stationary point is a global maximizer, with no constants involved.

Proposition K.1 (Exact optimality at square search rank)**.** *Let f be C^2 and concave on a neighborhood of $\mathcal{E}_N$, and let $F(V) = f(VV^\top)$ on $\mathcal{M}_{N,k}$. If $k \ge N$, $\operatorname{grad} F(V) = 0$ and $\operatorname{Hess} F(V) \preceq 0$, then $f(VV^\top) = \max_{Z \in \mathcal{E}_N} f(Z)$.*

Proof. The argument shows that the matrix $Q = \Lambda - A$ appearing in the factor Hessian (138) is positive semidefinite and annihilates V, after which concavity finishes the proof. Stationarity gives $QV = 0$, because the Riemannian gradient has rows $2((A-\Lambda)V)_i$. If V has full row rank, then $Q = 0$; otherwise $\operatorname{rank} V < N \le k$ supplies a nonzero $z \in \ker V$, and for any $u \in \mathbb{R}^N$ the matrix $U = uz^\top$ is tangent with $H = 0$, so the Hessian condition reads $-2\|z\|^2 u^\top Q u \le 0$. In either case $Q \succeq 0$. For any $Z \in \mathcal{E}_N$, the equal diagonals of Z and X together with $QV = 0$ give $\langle A, Z - X\rangle = -\langle Q, Z\rangle \le 0$, and concavity turns this into $f(Z) \le f(X)$, which is the claim. □

So at square search rank, second-order stationarity alone certifies global optimality on the elliptope, and no rate or curvature constant enters. The landscape bound and this proposition speak about points, and we now ask which points a descent method reaches in the limit. For the loss $L = -F$, write g_i for the sphere-projected gradient of row i and consider the normalized update

$$g_i = (I - v_i v_i^\top)\nabla_{v_i} L(V), \qquad (T_\alpha(V))_i = \frac{v_i - \alpha g_i}{\sqrt{1 + \alpha^2 \|g_i\|^2}}, \tag{139}$$

which moves each row against its projected gradient with step α and renormalizes it to the sphere.

Proposition K.2 (Analytic descent limits)**.** *If L is real analytic on $\mathcal{M}_{N,k}$, then for a sufficiently small fixed $\alpha > 0$, every trajectory of (139) converges to a critical point. For any initialization distribution with a density relative to manifold volume, the limiting Hessian is positive semidefinite almost surely.*

Proof. The argument has two parts: a descent estimate that makes the trajectory converge through the analytic Łojasiewicz inequality, and a stable-manifold argument that rules out convergence to points of negative curvature from almost every start. By compactness, $G_0 = \sup_V \|g(V)\|_F$ and the norm B_0 of the ambient derivative of g along unit tangent directions are finite, and differentiating (139) gives

$$\|DT_\alpha(V)[U] - U\|_F \le (\alpha B_0 + \alpha^2 G_0^2/2 + \alpha^2 G_0 B_0)\|U\|_F$$

for every tangent U. Once α is small enough that the coefficient is below one, $DT_\alpha(V)$ is injective between tangent spaces of equal dimension, so T_α is a local diffeomorphism of $\mathcal{M}_{N,k}$. The update is the row-normalization retraction applied to $V - \alpha g(V)$, and a uniform second-derivative bound along its tangent curves yields, after decreasing α further, the sufficient-decrease inequality $L(V) - L(T_\alpha V) \ge (\alpha/2)\|g(V)\|_F^2$. The displacement $\|T_\alpha V - V\|_F$ lies between a positive constant times $\alpha\|g(V)\|_F$ and $\alpha\|g(V)\|_F$, although only the upper bound is used below.

To invoke the convergence theorem of Absil et al. (2005), which is stated for analytic functions on Euclidean space, extend L zero-homogeneously off the manifold by $\tilde{L}(V) := L(\operatorname{rownorm} V)$, where rownorm divides every row by its Euclidean norm. This extension is real analytic on the open neighborhood of $\mathcal{M}_{N,k}$ on which no row vanishes. Because it is constant along rays through each row, its Euclidean gradient at a point of $\mathcal{M}_{N,k}$ is tangent and coincides with g, so that $\nabla\tilde{L} = g$ on the manifold and their theorem applies to the iterates as a sequence in $\mathbb{R}^{N\times k}$. Along that sequence, the sufficient-decrease inequality and the displacement bound combine to

$$\tilde{L}(V_t) - \tilde{L}(V_{t+1}) \ge \frac{\alpha}{2}\|g(V_t)\|_F^2 \ge \frac{1}{2}\|\nabla\tilde{L}(V_t)\|_F \, \|V_{t+1} - V_t\|_F,$$

which is the strong-descent hypothesis of their theorem with constant $\sigma = 1/2$. Whenever the loss fails to decrease, the gradient vanishes and the iterate is fixed. Together with the analytic Łojasiewicz gradient inequality for $\tilde{L}$, these two conditions give finite trajectory length and convergence to a single critical point.

At a critical point V_*, the differential $DT_\alpha(V_*)$ acts on the tangent space as $I - \alpha \operatorname{Hess} L(V_*)$, so a negative Hessian eigenvalue produces an eigenvalue of the map larger than one. Since T_α is a C^1 local diffeomorphism, the unstable-fixed-point theorem of Lee et al. (2019), which covers nonisolated fixed points, shows that almost every initial point avoids convergence to the set of such critical points. The limiting Hessian is therefore positive semidefinite almost surely. □

The proposition supplies a limit statement and no stopping time. Its analyticity hypothesis is stronger than the C^2 assumption of the landscape bound, although the polynomial penalized objectives in Appendix L satisfy it, whereas the positive-part penalty of Appendix L.9 does not.

L Applications and exact feasibility

The applications separate the established optimization problems from the additional guarantees supplied by the GD theorem and the landscape bounds. Table 3 gives the original formulations and closest prior guarantees, with theorem or equation locations where available. **R** denotes a recovered guarantee. **A** denotes a new prescribed-GD guarantee for an established optimization target. **E** denotes a new quantitative extension or application-specific bound relative to the cited comparator. Here "new" identifies the additional mathematical content relative to the cited results, specified in the last column; it does not claim a new optimization model or approximation ratio. The **A** rows inherit their algorithmic guarantee from Theorem 4.1. The **E** rows specialize our nonlinear landscape bounds; the CSP rounding and theta repair also require separate propositions. These specializations are consequences of the master results, rather than independent algorithmic frameworks. Recovering established guarantees within the same framework is part of the unification.

The finite-GD applications are proved in Appendices H and I. For the unit-row applications, Table 4 records the stored coordinates and the numerator $R + C/2$; the subsections below derive the constants and any subsequent feasibility or rounding step.

Table 2: Full application summary for the two main theorems; Table 1 gives a selection in the main text. The top group applies Theorem 4.1, using constant-step GD from $I_d/\sqrt{d}$ on a square factor. The bottom group applies Theorem 4.3(b). For unit-row factors of search rank k with $\operatorname{Hess} F \preceq \zeta I$, the gap to the relaxation optimum is at most $(R + C/2)/(k-1) + N\zeta/2$. The gap constant $R + C/2$ is listed first, with ρ the consistency penalty, B the nonconstant Fourier mass, and τ the contrastive temperature. Labels **R**, **A**, and **E** mark recovered guarantees, new prescribed-GD guarantees, and new quantitative extensions or application-specific bounds, respectively. Sources and the scope of each addition are detailed in Table 3.

Problem and precedent	Coordinates	Guarantee	Status / result
Corollaries of Theorem 4.1: square factor, finite-time gradient descent			
Bounded SDP feasibility (Hazan, 2008)	$(n+1)^2$	all constraint violations at most ε, or an infeasibility certificate, in $\widetilde{O}(d(L/\varepsilon)^2 \log q)$ steps	**A** Cor. 4.2
Nuclear-norm factorization; PSD sensing (Jaggi & Sulovský, 2010; Jaggi, 2013)	$(a+b)^2$; $(n+1)^2$	fitting objective within ε of the convex optimum over the norm ball	**A** Cors. H.2, H.6
Matrix completion (Candès & Recht, 2009)	$(a+b+1)^2$	observed entries exact, nuclear norm within ε of the minimum via norm-bound feasibility tests and entry repair	**A** Cor. H.7
Peng–Wei K-means SDP (Peng & Wei, 2007)	$(n+1)^2$	normalized constraint violations at most ε, objective threshold certified	**A** App. I
Corollaries of the landscape bound: unit-row factor of search rank k, gap constant $R + C/2$			
MAX-CUT (Mei et al., 2017; Goemans & Williamson, 1995)	nk	U_*: value at least $(1 - \frac{1}{k-1})U_* - \frac{n\zeta}{2}$, then Goemans–Williamson rounding	**R** App. L
Boolean MAX-a-CSP (Yau et al., 2024; Goemans & Williamson, 1994)	Nk, $N \le 1 + m2^a$	$2B + 32\rho 4^a$: penalized value, moment consistency, $3/4$ MAX-SAT rounding	**E + R** Cor. L.2
Level-r Lasserre ($2r \ge a$) (Lasserre, 2001b)	$k\sum_{j \le r}\binom{n}{j}$	$2B + 32\rho$: hierarchy value with mean squared consistency residual	**E** App. L
Lovász theta (Lovász, 1979)	$(n+1)k$	$1 + 21\rho/4$: certified interval for $\vartheta(G)/n$ after an explicit repair	**E** Prop. L.7
Contrastive embeddings (Khosla et al., 2020; Behnia & Thrampoulidis, 2024; Vock et al., 2026)	Nk	$4/\tau + 2/\tau^2$: in-sample loss within the bound of the convex optimum	**E** App. L
Two-layer Gram training (Nguyen & Hein, 2017; Levin et al., 2025)	$h(p+k)$	same bound with $\zeta = \gamma(\max_i \|z_i\|/\sigma_N(\Psi))^2$ at full feature rank and nonzero outputs	**E + R** Cor. 4.5

Table 3: Application provenance and contribution. **R**: recovered guarantee; **A**: new prescribed-GD guarantee for an established target; **E**: new quantitative extension or application-specific bound relative to the cited statement. Recovered guarantees form part of the unification. The last column credits the classical content and specifies each additional guarantee.

Application / status	Original formulation and closest prior result	Recovered content and precise addition here
Square factors: applications of the prescribed GD theorem		
Bounded SDP feasibility **A**	Hazan (2008, Theorems 1 and 3): bounded-trace optimization and smoothed feasibility. Zhang et al. (2023, Proposition 10): factor optimality certificate.	Approximate SDP solution and certification are established. Corollary 4.2 supplies the fixed polynomially regularized GD loss, start, step, and gradient-based certificate interval along a uniformly full-rank trajectory.
Nuclear-norm factorization **A**	Vandenberghe & Boyd (1996, Eq. (37)); Fazel et al. (2001, Lemma 1). Jaggi & Sulovský (2010, Lemma 1, Corollary 2, Theorem 3): exact trace lift and certified factor output.	The lift, norm-ball target, and additive fitting guarantee are recovered. Corollary H.2 realizes them through the prescribed ordinary-GD construction. No planted-matrix recovery is proved.
PSD sensing **A**	Hazan (2008, Theorem 1) and Jaggi (2013, Theorems 1–2, Section 4.3): smooth optimization over the PSD trace ball. Bhojanapalli et al. (2016, Theorem 5): factored GD with suitable initialization and matching search rank.	The convex least-squares target is established. Corollary H.6 supplies the prescribed GD realization for arbitrary sensing maps with known bounds. Recovery of an unknown matrix requires additional measurement assumptions.
Matrix completion **A**	Candès & Recht (2009, Eq. (1.5)): exact-observation nuclear-norm program. Cai et al. (2010, Theorems 3.1 and 4.2): singular-value-thresholding comparator.	The convex completion program is recovered. Corollary H.7 combines GD feasibility tests for a norm bound with entry repair, giving exact observed entries and an ε nuclear-norm gap. This uses several GD runs and gives no guarantee on unobserved entries.
Peng–Wei K-means SDP **A**	Peng & Wei (2007), Section 3.2 (Eq. (16) in the October 2005 author preprint): the SDP and its approximate solvability.	The relaxation is recovered. Appendix I gives its normalized residual system and prescribed-GD threshold certificate. A clustering approximation ratio requires an additional rounding argument.
Unit-row factors: recovered bounds and quantitative extensions		
MAX-CUT **R**	Mei et al. (2017, Theorem 2, Eq. (8)): the rank-dependent bound with $n\eta/2$. Goemans & Williamson (1995, Theorem 3.3): random-hyperplane rounding.	Both the displayed landscape bound and the rounding guarantee are recovered, including the Hessian-only hypothesis. This application adds no new approximation ratio or landscape theorem.

Continued on next page

Table 3 continued

Application / status	Original formulation and closest prior result	Recovered content and precise addition here
Boolean MAX-a-CSP / MAX-SAT **E + R**	Yau et al. (2024, Eq. (60), Theorem C.1, Lemma C.3, Corollary 3): penalized clause moments and robust rounding. Goemans & Williamson (1994, Theorem 5.2): $3/4$ marginal rounding.	The moment-penalty approach and $3/4$ ratio are recovered. The extension uses exact unit rows and every search rank, with Hessian-only numerator $2B + 32\rho 4^a$, direct rounding loss $3\sqrt{D}/2$, and the explicit rank prescription in Corollary L.2.
Level-r Lasserre **E**	Lasserre (2001a;b): polynomial moment hierarchy and exact Boolean SDP relaxation. Laurent (2003, Section 3.3, Eq. (22); Section 6.2.2, Eq. (33)): Boolean and sign-variable forms.	The hierarchy and full-level exactness are recovered. The extension is the normalized consistency penalty with numerator $2B + 32\rho$ and control of mean squared consistency residuals at fixed level r. Factor width does not increase the hierarchy level.
Lovász theta **E**	Lovász (1979, Theorems 4–5): theta representations. Laurent (2003, Section 6.1, Eq. (27)): theta body. Burer & Monteiro (2003, Section 4.1): low-rank theta optimization.	The relaxation and factorization are recovered. Proposition L.7 adds the explicit $1 + 21\rho/4$ penalty constant and feasibility-repair interval. This certifies the theta value; it does not round to an independent set.
Convex pairwise fitting **E**	Higham (2002a, Eqs. (1.1), (1.3)): weighted correlation fitting. Grubišić & Pietersz (2007, Eqs. (1.1)–(1.2), Section 3): rank-constrained unit-row formulation.	The squared-loss formulation and factor geometry are recovered. The extension is the Hessian-only loss-gap bound with numerator $2L_1 + 2L_2$ for the stated smooth convex pairwise family.
Contrastive embeddings **E**	van den Oord et al. (2018, Eq. (4)): InfoNCE; Chen et al. (2020, Eq. (1)) and Khosla et al. (2020, Eq. (2)): normalized losses. Behnia & Thrampoulidis (2024, Theorem 1): local-to-global result on norm balls under their dimension and temperature assumptions; Vock et al. (2026, Theorem 3.13): global SupCon geometry for every $\tau > 0$ when the number of classes does not exceed the embedding dimension.	The losses, convex Gram viewpoint, and exact-globality results under structured labels are established. Our addition is the approximate-Hessian gap with numerator $4/\tau + 2/\tau^2$ for arbitrary fixed J_i, q_i and every search rank $k \geq 2$. Allowing every $\tau > 0$ alone is not new; this bound does not reproduce the cited exact optimum geometry.
Penalized K-means SDP **E**	Peng & Wei (2007): the same classical SDP. The linear unit-row landscape argument is from Mei et al. (2017, Theorem 2).	Proposition L.8 supplies the $2n$-row lift and its $n + 32\rho$ penalized gap and average-feasibility calculation. The inspected output can violate the SDP constraints; no clustering rounding guarantee is proved.
Two-layer Gram training **E + R**	Nguyen & Hein (2017, Lemma 3.5): full-feature-rank mechanism. Levin et al. (2025, Proposition 2.12): exact criticality transfer through a submersion.	Exact criticality transfer is recovered. Corollary 4.5 adds the tolerance $\gamma(\max_i \|z_i\|/\sigma_N(\Psi))^2$ combined with the Gram gap bound. Full feature rank and nonzero outputs are assumed at the inspected point.

Continued on next page

Table 3 continued

Application / status	Original formulation and closest prior result	Recovered content and precise addition here
Two-layer prediction **R**	Nguyen & Hein (2017, Lemma 3.5): stationary output layer with full feature rank. Haeffele & Vidal (2015, Theorem 15), later Haeffele & Vidal (2017, Theorem 1): zero-neuron local-minimum certificate.	Both certificates are recovered specializations. The second requires positive homogeneity, matched regularization, and an unused neuron at a local minimum. They provide no guarantee that joint training reaches such a point.

Table 4: Unit-row applications on $\mathcal{M}_{N,k}$, with Nk stored coordinates and intrinsic dimension $N(k-1)$. Here n and m count variables and constraints, a is the maximum arity, U_* is the cut SDP value, ρ is the consistency penalty, and B is the weighted nonconstant Fourier coefficient mass; pairwise losses obey $|\ell'| \le L_1$ and $0 \le \ell'' \le L_2$, and the contrastive temperature is $\tau > 0$. The Lasserre row uses $2r \ge a$ and representative moments, the clustering row uses normalized squared distances and a feasibility penalty, and the final row records separate neural certificates with hidden width h and input dimension p. The status labels and exact source comparisons are in Table 3. The subsections below define each objective and derive its constants.

Problem	Rows N / state size	Numerator $R + C/2$	Guarantee before rounding
Normalized MAX-CUT (**R**)	n	U_*	$F \ge (1 - 1/(k-1))U_* - n\eta/2$
Boolean MAX-a-CSP (**E + R**)	$\left\lvert\bigcup_C 2^C\right\rvert \le 1 + m2^a$	$2B + 32\rho 4^a$	Penalized SDP gap and average consistency control
Level-r Boolean Lasserre (**E**)	$N_r = \sum_{j=0}^{r} \binom{n}{j}$	$2B + 32\rho$	Penalized hierarchy gap with mean squared consistency residuals
Lovász theta (**E**)	$n + 1$	$1 + 21\rho/4$	Finite certified interval for $\vartheta(G)/n$ after feasibility repair
Convex pairwise fitting (**E**)	Number of items	$2L_1 + 2L_2$	Average loss within $(2L_1 + 2L_2)/(k-1) + N\eta/2$ of the convex optimum
Contrastive embeddings (**E**)	Number of samples	$4/\tau + 2/\tau^2$	Average training loss within the stated additive bound
K-means SDP (**E**)	$2n$ for n data points	$n + 32\rho$	Penalized SDP gap and average feasibility control
Two-layer Gram training (**E + R**)	N samples $h(p+k)$ parameters	As for the Gram loss	Same bound via Corollary 4.5, with full feature rank and nonzero head outputs
Two-layer prediction losses (**R**)	$h(p+1)$ parameters	Separate certificates	Global optimum for convex prediction losses at a stationary output layer with full feature rank, or a local minimum with an unused neuron and matched regularization

L.1 MAX-CUT

The simplest instance is the cut objective, for which the nonlinear curvature vanishes. Let $w_{ij} \geq 0$ be edge weights normalized so that $\sum_{i<j} w_{ij} = 1$, and set $f(X) = \frac{1}{2}\sum_{i<j} w_{ij}(1 - X_{ij})$, the normalized cut value of a correlation matrix. Over the elliptope this linear function attains its minimum zero at $X = \mathbf{1}\mathbf{1}^\top$ and its maximum at the cut SDP value U_*, so that $R(\nabla f) = U_*$. Since f is linear, $C = 0$, and the landscape bound recovers Mei et al. (2017, Theorem 2), as recorded in Table 4. Random-hyperplane rounding of any elliptope point X has expected cut value at least $\alpha_{\mathrm{GW}} f(X)$ with the classical $\alpha_{\mathrm{GW}} \simeq 0.87856$ (Goemans & Williamson, 1995), so the additive loss $n\eta/2$ is multiplied by the same factor after rounding.

L.2 Boolean CSP moments

Boolean constraint satisfaction replaces the single cut objective by a weighted predicate score written in clause-local moments, and the curvature now comes from a consistency penalty rather than from the objective. Let $P_C \colon \{-1,+1\}^C \to [0,1]$ be predicates, each on a clause C of at most a variables, with positive weights w_C summing to one. The moment index set $\mathcal{I} = \bigcup_C 2^C$ consists of all subsets of clauses, giving $N = |\mathcal{I}|$ rows, and the moment of a subset T is read from the anchor row as $y_T = X_{\varnothing,T}$. Expanding each predicate in its Fourier series and evaluating it on these moments gives the score and the nonconstant coefficient mass,

$$P_C(x) = \sum_{T \subseteq C} \widehat{P}_C(T) \prod_{i \in T} x_i, \quad S(X) = \sum_C w_C \sum_{T \subseteq C} \widehat{P}_C(T) y_T, \quad B = \sum_C w_C \sum_{\varnothing \neq T \subseteq C} |\widehat{P}_C(T)|,$$

and with $a_0 = \sum_C w_C \widehat{P}_C(\varnothing)$ for the constant term, the score obeys $S(X) \leq U := a_0 + B$ on the whole elliptope. Consistency of the clause-local moments is measured over ordered pairs of subsets of each clause. The residuals, their weighted square sum, and the penalized objective with penalty weight ρ are

$$r_{C,A,B} = X_{A,B} - X_{\varnothing, A \triangle B}, \qquad D(X) = \sum_C w_C \sum_{A,B \subseteq C} r^2_{C,A,B}, \qquad f_\rho = S - \rho D,$$

so that $D = 0$ is exact clause moment consistency. For a consistent Gram factor V with rows v_A indexed by $A \in \mathcal{I}$, the quantities

$$\mu_C(x) = \left\| 2^{-|C|} \sum_{A \subseteq C} \left(\prod_{i \in A} x_i \right) v_A \right\|^2, \qquad x \in \{-1,+1\}^C,$$

are nonnegative and sum to one by character orthogonality, so each clause carries a probability law. Consistency makes the moments of that law $\mathbb{E}_{\mu_C} \prod_{i \in T} x_i = y_T$, and the score S is then the expected predicate value under these local laws. Since every Boolean assignment lifts to a consistent factor, $0 \leq \mathrm{OPT} \leq c_* := \max_{\mathcal{E}_N, D=0} S \leq 1$, where c_* is the value of the consistent relaxation.

To apply the landscape bound we need the linear range and the curvature of f_ρ. Each residual has off-diagonal coefficient sum at most two and magnitude at most two, and a clause of arity at most a has at most 4^a ordered pairs of subsets. The gradient of f_ρ therefore has coefficient sum at most $B + 8\rho 4^a$, and $R(\nabla f_\rho) \leq 2B + 16\rho 4^a$. For the curvature, the entry bound (137) and the L^2 triangle inequality give $\mathbb{E}(\mathrm{d}r[H])^2 \leq 16/k$ for every residual, and since $\nabla^2 f_\rho[H,H] = -2\rho \sum_C w_C \sum_{A,B} (\mathrm{d}r[H])^2$, we may take $C = 32\rho 4^a$. With the resulting defect and the penalized optimum,

$$\Delta = \frac{2B + 32\rho 4^a}{k-1} + \frac{N\eta}{2}, \qquad c_\rho = \max_{X \in \mathcal{E}_N} f_\rho(X),$$

the landscape bound yields

$$0 \leq c_\rho - f_\rho(X) \leq \Delta, \qquad S(X) \geq c_* - \Delta, \qquad D(X) \leq \frac{U + \Delta}{\rho}, \tag{140}$$

where the last inequality uses $c_\rho \geq c_* \geq 0$ and $\rho D = S - f_\rho$. For MAX-SAT, $B \leq 1$ and the simpler bound $U = 2$ is valid. When the Hessian bound has been verified at the current point, the same inequalities give the upper certificate $\mathrm{OPT} \leq c_* \leq f_\rho(X) + \Delta$, even if that point has a nonzero consistency residual.

L.3 Direct rounding with inconsistent moments

The preceding CSP bounds control the weighted sum of squared consistency residuals, so a rounding scheme that needs exact clause laws would first have to repair the moments. This section shows that the classical MAX-SAT marginal rounding of Goemans & Williamson (1994) survives a finite defect without any repair. We use the clause-local MAX-SAT moments and their unnormalized ordered-pair residuals, and for a clause C of size d we write $D_C = \sum_{A,B\subseteq C} r_{C,A,B}^2$ for its sum of squared consistency residuals.

Proposition L.1 (Marginal rounding with inconsistent moments)**.** *Set $p_i = (1 + X_{\varnothing,\{i\}})/2$. Independent rounding with probability p_i of setting $x_i = +1$ has expected weighted score at least $\alpha_a(S - \frac{3}{2}\sqrt{D})$, where $\alpha_a = 1 - (1 - 1/a)^a$. Independent rounding with probability $1 - 4^{-p_i}$ instead has expected score at least $\frac{3}{4}(S - \frac{3}{2}\sqrt{D})$. Both bounds can be derandomized by conditional expectations.*

Proof. Three steps are needed: the singleton marginals read from the factor are compared with the marginals of the local law of each clause, the discrepancy is bounded using these squared residuals, and the classical rounding analysis is run on the corrected marginals. For every clause define $\mu_C(x) = \|2^{-d}\sum_{A\subseteq C} x_A v_A\|^2$ for $x \in \{-1,+1\}^C$, where $x_A = \prod_{i\in A} x_i$; character orthogonality shows that this is a probability law. Writing $\nu_C(x) = 2^{-d}\sum_{T\subseteq C} x_T y_T$ for the law that exact consistency would produce, expansion gives

$$\mu_C(x) - \nu_C(x) = 2^{-2d} \sum_{A,B\subseteq C} x_{A\triangle B} r_{C,A,B},$$

so Cauchy–Schwarz yields $|\mu_C(x) - \nu_C(x)| \le 2^{-d}\sqrt{D_C}$. The singleton marginals of μ_C satisfy

$$\mu_C(x_i = s) - \frac{1 + s y_i}{2} = s 2^{-d-1} \sum_{A\triangle B = \{i\}} r_{C,A,B},$$

whose absolute value is at most $2^{-d/2-1}\sqrt{D_C}$ by Cauchy–Schwarz over the 2^d pairs with $A\triangle B = \{i\}$. For the unique violating assignment x^*, the canonical score is $s_C = 1 - \nu_C(x^*)$, so $s_C \le 1 + 2^{-d}\sqrt{D_C}$.

Let p_ℓ be the probability that literal ℓ is true under the untransformed product measure. The law μ_C puts mass $1 - \mu_C(x^*)$ on assignments satisfying C, so the union bound under μ_C together with the two discrepancy estimates gives

$$\sum_{\ell\in C} p_\ell \ge s_C - e_C, \qquad e_C = (2^{-d} + d2^{-d/2-1})\sqrt{D_C} \le \sqrt{D_C},$$

where the last scalar inequality follows by checking $d = 1, 2, 3$ and observing that $d2^{-d/2-1}$ decreases for integer $d \ge 3$. Put $t_C = \min\{1, \max\{0, s_C - e_C\}\}$; then $t_C \le \sum_{\ell\in C} p_\ell$, and $t_C \ge s_C - \frac{3}{2}\sqrt{D_C}$ because clipping at zero can only increase $s_C - e_C$, while clipping at one loses at most $2^{-d}\sqrt{D_C} \le \frac{1}{2}\sqrt{D_C}$.

For the original product measure, the arithmetic–geometric mean inequality gives clause success probability at least $1 - (1 - t_C/d)^d$, and since this function of t_C is concave on $[0, 1]$ and vanishes at zero, it is at least $[1 - (1 - 1/d)^d]t_C \ge \alpha_a t_C$.

For the transformed rounding, a positive literal is false with probability $4^{-p_i} = 4^{-p_\ell}$, whereas a negative literal is false with probability $1 - 4^{-p_i}$, and for $u = 4^{-p_i} \in [1/4, 1]$,

$$4^{p_i - 1} - (1 - 4^{-p_i}) = \frac{1}{4u} - 1 + u = \frac{(1 - 2u)^2}{4u} \ge 0,$$

so the failure probability of a negative literal is also at most 4^{-p_ℓ}. Independence then bounds the clause failure probability by $4^{-\sum p_\ell} \le 4^{-t_C}$, and concavity gives $1 - 4^{-t_C} \ge (3/4)t_C$. Summing either guarantee over clauses and using $\sum_C w_C\sqrt{D_C} \le \sqrt{\sum_C w_C D_C} = \sqrt{D}$ proves the two expected-score bounds. For the derandomization, expose one variable at a time and select the value with the larger conditional expected score, which never decreases the expectation, so the final assignment attains at least the expected score. □

The proposition converts an average consistency defect D into a $\frac{3}{2}\sqrt{D}$ loss of score, and combining it with the CSP landscape bound gives an explicit search rank for the classical three-quarter guarantee.

Corollary L.2 (Explicit rank for the classical three-quarter factor). *For $0 < \varepsilon \le 1$, choose*

$$\rho = 27\varepsilon^{-2}, \qquad k - 1 \ge 12\varepsilon^{-1} + 5184\, 4^a \varepsilon^{-3}, \qquad \eta \le \frac{\varepsilon}{3N}.$$

Every factor obeying the loss Hessian bound $\mathrm{Hess}(-f_\rho \circ \mathrm{Gram}) \succeq -\eta I$ rounds to score at least $\frac{3}{4}(c_* - \varepsilon)$ in expectation. Conditional expectations yield an assignment with at least that score.

Proof. The stated choices give $\Delta = (2 + 32\rho 4^a)/(k-1) + N\eta/2 \le \varepsilon/3$, so Equation (140) gives $S \ge c_* - \varepsilon/3$ and $D \le 7\varepsilon^2/81$, and therefore $S - \frac{3}{2}\sqrt{D} \ge c_* - (1/3 + \sqrt{7}/6)\varepsilon \ge c_* - \varepsilon$. Proposition L.1 now gives the expected score and its derandomization. □

Because $c_* \ge \mathrm{OPT}$, the corollary is a classical three-quarter approximation up to the stated additive error, obtained at a finite penalty from a Hessian bound verified at a finite iterate. It does not bound the number of iterations needed to reach that Hessian bound, and since the residual normalization here is clause-local, the same constants must not be transferred unchanged to the normalized Lasserre mean squared consistency residual D_r, defined in the next subsection.

L.4 The complete Boolean Lasserre moment constraints

The clause-local moments used for CSPs and rounding above involve only the subsets of single clauses, and the natural strengthening indexes rows by all small subsets of variables, which is the Lasserre hierarchy. Fix a level $\lceil a/2 \rceil \le r \le n$ and use the index set $\mathcal{I}_r = \{A \subseteq [n] \mid |A| \le r\}$ with $N_r = \sum_{j=0}^{r} \binom{n}{j}$ rows. For each symmetric difference T realized by a pair in $\mathcal{I}_r^2$, choose one representative pair (A_T, B_T) with $A_T \triangle B_T = T$, and measure consistency by the mean squared consistency residual

$$D_r(X) = \frac{1}{N_r^2} \sum_{A,B \in \mathcal{I}_r} (X_{A,B} - X_{A_{A\triangle B}, B_{A\triangle B}})^2. \tag{141}$$

The feasible condition $D_r = 0$ says that $X_{A,B}$ depends only on $A \triangle B$, or equivalently that there are moments y_T for $|T| \le \min(2r, n)$ with $X_{A,B} = y_{A\triangle B}$, and together with $X \succeq 0$ and $y_\varnothing = 1$ this is the Boolean level-r Lasserre moment relaxation (Lasserre, 2001b; Laurent, 2003). For the objective we define $y_T = X_{A_T, B_T}$ through the chosen representative, which is available for every predicate term because $2r \ge a$. When $|T| \le r$ we choose $(A_T, B_T) = (\varnothing, T)$ so as to recover the canonical moment convention, and at exact consistency the objective does not depend on these choices. Boolean lifts still show $c_*^{(r)} \ge \mathrm{OPT} \ge 0$, although for $r < a$ the relaxation can have $c_*^{(r)} > 1$ because it need not supply full clause probability laws; the elliptope bound $S \le a_0 + B$ remains sufficient for the penalty argument.

Because of the normalization in (141), $0 \le D_r \le 4$, and the same coefficient calculation as for the clause-local model yields $R \le 2B + 16\rho$ and $C \le 32\rho$ for $f_\rho = S - \rho D_r$, so that (140) holds with $\Delta_r = (2B + 32\rho)/(k-1) + N_r\eta/2$. Only the mean squared consistency residual is controlled at finite penalty, and a single constraint can require stronger accuracy as N_r grows.

Since principal submatrices preserve the moment constraints, the hierarchy values decrease, $c_*^{(n)} \le \cdots \le c_*^{(\lceil a/2 \rceil)}$, and at the top level they reach the discrete optimum. For $r = n$, character orthogonality applied to the full group moment matrix gives the law $\mu(x) = 2^{-n} \sum_{T \subseteq [n]} y_T \prod_{i \in T} x_i$, which equals $\|2^{-n} \sum_A (\prod_{i \in A} x_i) v_A\|^2$ and is therefore nonnegative, sums to one, and reproduces every moment, so that $c_*^{(n)} = \mathrm{OPT}$. This stronger comparison value does not by itself specify a rounding algorithm or a finite running time.

Why all difference classes are necessary. The mean squared consistency residual (141) compares every pair with a representative of its difference class, including classes with $|A \triangle B| > r$ that have no anchor entry $X_{\varnothing,T}$, and the following example shows that the truncated equalities alone would be too weak. For $n = 3$ and $r = 2$, start from the identity matrix indexed by $\varnothing, 1, 2, 3, 12, 13, 23$ and set

only $X_{1,23} = X_{23,1} = 1/4$ off the diagonal. Its least eigenvalue is $3/4$, so it belongs to the elliptope, and every equality comparing to $X_{\varnothing,A\triangle B}$ with $|A\triangle B| \le r$ holds, yet the entries $X_{1,23}, X_{2,13}, X_{3,12}$ differ, so those truncated equalities do not define the Lasserre relaxation.

L.5 Penalty limits

The finite-penalty bounds above control aggregate squared consistency residuals, and sending the penalty to infinity recovers the exact relaxation in the limit. For either moment model, fix the instance and suppose that $\rho_j \to \infty$ and $f_{\rho_j}(X_j) \ge c_* - \delta_j$ with $\delta_j \to 0$. Then $D(X_j) \le (U + \delta_j)/\rho_j \to 0$, so every accumulation point is feasible, and since $S(X_j) \ge c_* - \delta_j$, every such point has score c_*. Compactness therefore implies

$$\operatorname{dist}_F(X_j, \mathcal{X}_*) \to 0, \quad S(X_j) \to c_*, \quad f_{\rho_j}(X_j) \to c_*, \quad \rho_j D(X_j) \to 0,$$

where $\mathcal{X}_*$ is the optimal set of the exact relaxation, and the last two limits follow by squeezing $f_{\rho_j}(X_j)$ between $c_* - \delta_j$ and $S(X_j)$. Proposition K.2 and a countable intersection of probability-one events give these limits for separately converged runs with $\rho_j/(k_j - 1) \to 0$, without any independence between the runs, and Proposition K.1 alternatively permits a fixed square search rank $k \ge N$.

L.6 Uniform affine feasibility and refutation

The CSP and Lasserre bounds of Appendix L control weighted sums or averages of squared consistency residuals, and a single constraint may still be badly violated. This subsection returns to the running feasibility example and shows that the landscape bound of Theorem 4.3(b) also controls every constraint of an affine system on the elliptope at once, provided the maximum constraint violation is first smoothed by log-sum-exp. The resulting bound is then turned into a refutation test. Here $\mathcal{E}_N = \{X \in \mathbb{R}^{N\times N} \mid X = X^\top,\ X \succeq 0,\ X_{ii} = 1\}$ is the elliptope and $\mathcal{M}_{N,k} = (\mathbb{S}^{k-1})^N$ is the manifold of unit-row factors with search rank k. The map $\operatorname{Gram}(V) = VV^\top$ sends a factor to the elliptope, and all factor Hessians use the product-sphere metric. For $M \ge 1$, consider the affine inequalities

$$a_\ell(X) = b_\ell + \sum_{i<j} c_{\ell,ij} X_{ij} \le 0, \qquad \sum_{i<j} |c_{\ell,ij}| \le L, \qquad 1 \le \ell \le M,$$

where the coefficient bound $L > 0$ bounds every constraint, and where diagonal terms have been absorbed into b_ℓ because $X_{ii} = 1$ on the elliptope. The maximum violation, its minimum over the elliptope, and the log-sum-exp smoothing at temperature $\tau > 0$ are

$$v(X) = \max\{0, a_1(X), \dots, a_M(X)\}, \qquad \gamma = \min_{X\in\mathcal{E}_N} v(X),$$

$$E_\tau(X) = \tau \log\left(1 + \sum_{\ell=1}^{M} e^{a_\ell(X)/\tau}\right),$$

where compactness guarantees that γ is attained, and the affine system is feasible if and only if $\gamma = 0$.

The first result keeps the temperature free, so that the trade-off between smoothing error and curvature is visible before it is optimized.

Proposition L.3 (Arbitrary smoothing temperature)**.** *For $k \ge 2$, $\eta \ge 0$ and $V \in \mathcal{M}_{N,k}$, suppose* $\operatorname{Hess}(E_\tau \circ \operatorname{Gram})(V) \succeq -\eta I$. *Then*

$$v(VV^\top) \le \gamma + \frac{2L + 2L^2/\tau}{k-1} + \tau \log(M+1) + \frac{N\eta}{2}. \tag{142}$$

Proof. We apply the landscape bound to $f = -E_\tau$, computing its linear range from the coefficient bound and its probe curvature from the log-sum-exp variance formula, and then remove the smoothing with the elementary softmax inequalities. Put $a_0 = 0$ and let $p_\ell = e^{a_\ell(X)/\tau} / \sum_{j=0}^{M} e^{a_j(X)/\tau}$ be the softmax weights. The function $f = -E_\tau$ is analytic and concave on the symmetric matrices. Its differential is a negative sub-convex combination of the differentials $\mathrm{d}a_\ell$ (the weights $p_1, \dots, p_M$ sum to $1 - p_0 \le 1$), so the sum of its absolute off-diagonal coefficients is at most L. Since two elliptope entries differ by at most two, the linear range in Theorem 4.3(b) satisfies $R(\nabla f(X)) \le 2L$.

For the curvature we use the Gaussian probe of that theorem: write $X = VV^\top$, factor a comparator as $X_* = YY^\top$, and set $u_i = (I - v_i v_i^\top)Gy_i$ and $H = VU^\top + UV^\top$ for independent Gaussian entries of G with variance $1/k$, where U has rows u_i. For $a = X_{ij}$ and $b = (X_*)_{ij}$, the two summands of H_{ij} each have variance $(1-a^2)/k$ and covariance $-ab(1-a^2)/k$, so that

$$\mathbb{E}H_{ij}^2 = \frac{2}{k}(1 - X_{ij}^2)(1 - X_{ij}(X_*)_{ij}) \le \frac{4}{k},$$

and the triangle inequality in L^2 gives

$$\big(\mathbb{E}(\mathrm{d}a_\ell[H])^2\big)^{1/2} \le \sum_{i<j} |c_{\ell,ij}|(\mathbb{E}H_{ij}^2)^{1/2} \le \frac{2L}{\sqrt{k}}.$$

Because the log-sum-exp Hessian is a covariance,

$$-\nabla^2 f(X)[H,H] = \frac{1}{\tau}\operatorname{Var}_{\ell\sim p}(\mathrm{d}a_\ell[H]) \le \frac{1}{\tau}\sum_{\ell=0}^{M} p_\ell(\mathrm{d}a_\ell[H])^2,$$

and taking expectations gives $-k\mathbb{E}\nabla^2 f(X)[H,H] \le 4L^2/\tau$. The factor Hessian assumption reads $\mathrm{Hess}(f \circ \mathrm{Gram})(V) \preceq \eta I$ in the sign convention of the theorem, so applying it with a minimizer X_* of v as comparator yields

$$E_\tau(X) - E_\tau(X_*) \le \frac{2L + 2L^2/\tau}{k-1} + \frac{N\eta}{2}.$$

Finally, the elementary softmax inequalities $v(Z) \le E_\tau(Z) \le v(Z) + \tau\log(M+1)$ hold for every Z, and using them at X and at X_* turns the last display into (142). □

The two temperature-dependent terms of the bound pull in opposite directions, and balancing them gives the final statement, which we record with all hypotheses restated so that it can be quoted on its own.

Corollary L.4 (Uniform feasibility and refutation)**.** *Let $M \ge 1$ affine constraints be $a_\ell(X) = b_\ell + \sum_{i<j} c_{\ell,ij}X_{ij} \le 0$, with $\sum_{i<j}|c_{\ell,ij}| \le L$ and $L > 0$. Put $v(X) = \max\{0, a_1(X), \dots, a_M(X)\}$, $\gamma = \min_{X\in\mathcal{E}_N} v(X)$ and $E_\tau(X) = \tau\log(1 + \sum_\ell e^{a_\ell(X)/\tau})$. For $k \ge 2$ and $\eta \ge 0$, choose $\tau = L\sqrt{2/((k-1)\log(M+1))}$. If $\mathrm{Hess}(E_\tau \circ \mathrm{Gram})(V) \succeq -\eta I$, then*

$$\begin{aligned} v(VV^\top) &\le \gamma + B_{k,\eta}, \\ B_{k,\eta} &= \frac{2L}{k-1} + 2\sqrt{2}L\sqrt{\frac{\log(M+1)}{k-1}} + \frac{N\eta}{2}. \end{aligned} \tag{143}$$

Proof of Corollary L.4. The terms of (142) that depend on τ are $2L^2/((k-1)\tau) + \tau\log(M+1)$, which are minimized at $\tau = L\sqrt{2/((k-1)\log(M+1))}$, where their sum equals $2\sqrt{2}L\sqrt{\log(M+1)/(k-1)}$; substituting this value into (142) gives (143). □

The corollary says that a factor whose loss Hessian has no eigenvalue below $-\eta$ violates no constraint by more than $B_{k,\eta}$ beyond the best possible violation γ. When the system is feasible, $\gamma = 0$ and the bound controls every constraint simultaneously, so a search rank $k = O(L/\varepsilon + L^2\log(M+1)/\varepsilon^2)$ together with $\eta = O(\varepsilon/N)$ gives uniform $O(\varepsilon)$ violation. Read in the other direction, at a point where the Hessian bound has been verified the inequality supplies the lower bound $\gamma \ge v(VV^\top) - B_{k,\eta}$, so that $v(VV^\top) > B_{k,\eta}$ proves infeasibility, whereas positive violation at a candidate point or a small gradient on their own prove nothing.

L.6.1 The SAT specialization

The refutation test applies to the clause-local SAT relaxation once its constraints are written as an affine system with a coefficient bound. Consider a CNF formula whose nonempty clauses contain distinct variables, after repeated literals and tautological clauses have been removed; an empty clause already proves unsatisfiability. Use the moment indices $\mathcal{I} = \bigcup_C 2^C$ with $N = |\mathcal{I}|$ rows, and for a

literal on variable i in clause C write $\sigma_{C,i} \in \{-1,+1\}$ for its satisfying sign and $\sigma_{C,T} = \prod_{i\in T} \sigma_{C,i}$ for the sign of a subset. For $X \in \mathcal{E}_N$, set $y_T = X_{\varnothing,T}$ and let

$$s_C(X) = 1 - 2^{-|C|} \sum_{T\subseteq C} (-1)^{|T|} \sigma_{C,T} y_T$$

be the clause score, which at exact consistency is the probability under the clause-local law that C is satisfied. The affine system consists of the two-sided consistency constraints and the requirement that every clause be satisfied,

$$\begin{aligned} \frac{1}{2}\big(X_{A,B} - X_{\varnothing,A\triangle B}\big) &\le 0, \\ -\frac{1}{2}\big(X_{A,B} - X_{\varnothing,A\triangle B}\big) &\le 0 \quad (A,B \subseteq C), \\ 1 - s_C(X) &\le 0, \end{aligned}$$

whose off-diagonal coefficient sums are at most one, so that $L = 1$ is a valid coefficient bound. For any satisfying assignment $x \in \{-1,+1\}^n$, the Boolean lift $v_A = (\prod_{i\in A} x_i)e$ with a fixed unit vector e has exact moment consistency and $s_C = 1$ for every clause, so satisfiability implies $\gamma = 0$. A verified Hessian bound together with $v(VV^\top) > B_{k,\eta}$ therefore proves unsatisfiability. An unsatisfiable formula can still have a feasible clause-local relaxation, in which case this test cannot refute it, and larger moment relaxations can be handled by the same affine theorem once their complete constraint system and coefficient bound are specified.

L.7 Lovász theta and a finite certified interval

The affine feasibility bound of Appendix L.6 treats constraints as a system to be satisfied or refuted, whereas the Lovász theta function asks for the optimum of a linear score subject to equality constraints on the elliptope. This subsection penalizes those constraints, applies the landscape bound to the penalized score, and then repairs the constraint violation explicitly, so that both endpoints of an interval containing $\vartheta(G)$ are read off one factor. Let $G = ([n], E)$ be a simple graph with $n \ge 2$ vertices and $m = |E| > 0$ edges, and write $\alpha(G)$ for its maximum independent-set size; the edgeless case has $\alpha(G) = \vartheta(G) = n$ and needs no optimization. We use the standard theta-body formulation (Gaar & Rendl, 2020, Section 2.2),

$$\begin{aligned} \vartheta(G) = \max & \sum_{i=1}^{n} \|u_i\|^2 \\ \text{subject to} \quad & \|u_0\| = 1, \quad \langle u_0, u_i\rangle = \|u_i\|^2 \quad (1 \le i \le n), \\ & \langle u_i, u_j\rangle = 0 \quad (ij \in E), \end{aligned} \tag{144}$$

in which the maximum ranges over finite-dimensional vector families whose Gram matrix makes the problem a semidefinite program.

To place this program on the elliptope, let $\mathcal{E}_{n+1} = \{X = X^\top \succeq 0 \mid X_{ii} = 1\}$ have rows and columns indexed by $0, 1, \dots, n$, where index zero is an optimized unit anchor rather than a fixed vector. For $X \in \mathcal{E}_{n+1}$ we define the normalized score, the edge residuals, and the theta feasible set

$$\begin{aligned} S_G(X) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1 + X_{0i}}{2}, \qquad r_{ij}(X) = \frac{1 + X_{0i} + X_{0j} + X_{ij}}{4} \quad (ij \in E), \\ \mathcal{T}_G &= \{X \in \mathcal{E}_{n+1} \mid r_{ij}(X) = 0 \ (ij \in E)\}. \end{aligned} \tag{145}$$

The lemma below checks that this feasible set reproduces the theta body, so that every later bound on S_G over $\mathcal{T}_G$ is a bound on $\vartheta(G)/n$.

Lemma L.5 (Theta as an affine slice of the elliptope)**.** *The set $\mathcal{T}_G$ is nonempty and compact, and*

$$\theta_* := \max_{X\in\mathcal{T}_G} S_G(X) = \frac{\vartheta(G)}{n}, \qquad \frac{\alpha(G)}{n} \le \theta_* \le 1.$$

Proof. The two formulations are related by the affine change of vectors $u_i = (v_0 + v_i)/2$, which we check in both directions. Factor $X \in \mathcal{T}_G$ as $X = VV^\top$ with unit rows v_i and set $u_0 = v_0$ and $u_i = (v_0 + v_i)/2$; then

$$\langle u_0, u_i\rangle = \|u_i\|^2 = (1 + X_{0i})/2, \qquad \langle u_i, u_j\rangle = r_{ij}(X) = 0 \quad (ij \in E),$$

which is a theta-body representation with score $nS_G(X)$. Conversely, for vectors satisfying (144), put $v_0 = u_0$ and $v_i = 2u_i - u_0$, whose squared norms are $4\|u_i\|^2 - 4\langle u_0, u_i\rangle + 1 = 1$, so that their Gram matrix lies in $\mathcal{T}_G$ and preserves the score; the two optima therefore agree. For an independent set I, taking $v_0 = e$, $v_i = e$ for $i \in I$, and $v_i = -e$ otherwise makes every residual vanish with $nS_G = |I|$, which proves the lower bound, and the empty set shows that $\mathcal{T}_G$ is nonempty. The upper bound follows from $0 \le S_G(X) \le 1$ on the whole elliptope, and $\mathcal{T}_G$ is compact as a closed subset of a compact set. □

Rather than enforce the edge equations, we penalize them. For a penalty weight $\rho > 0$, define the mean squared edge residual, the penalized score, its factor version, and the penalized optimum

$$\begin{aligned} D_G(X) &= \frac{1}{m}\sum_{ij\in E} r_{ij}(X)^2, \qquad f_\rho^G(X) = S_G(X) - \rho D_G(X), \\ F_\rho^G(V) &= f_\rho^G(VV^\top), \qquad c_\rho^G = \max_{X\in\mathcal{E}_{n+1}} f_\rho^G(X). \end{aligned} \tag{146}$$

The function f_ρ^G is concave and polynomial in X, its factor domain is $\mathcal{M}_{n+1,k} = (\mathbb{S}^{k-1})^{n+1}$, and all Hessians below use the Frobenius metric of this manifold.

Proposition L.6 (Landscape bound and upper certificate)**.** *For $k \ge 2$, $\eta \ge 0$ and $V \in \mathcal{M}_{n+1,k}$, suppose* $\operatorname{Hess} F_\rho^G(V) \preceq \eta I$. *Define*

$$\Delta_G = \frac{1 + 21\rho/4}{k-1} + \frac{(n+1)\eta}{2}.$$

With $X = VV^\top$, one has

$$\begin{aligned} 0 \le c_\rho^G - F_\rho^G(V) \le \Delta_G, \\ S_G(X) \ge \theta_* - \Delta_G, \qquad D_G(X) \le \frac{1+\Delta_G}{\rho}. \end{aligned} \tag{147}$$

In particular, $\alpha(G)/n \le \vartheta(G)/n \le F_\rho^G(V) + \Delta_G$.

Proof. Both constants of Theorem 4.3(b) for f_ρ^G come from counting coefficients, after which the penalized gap is converted into bounds on the score and the residual. The sum of the absolute unordered off-diagonal coefficients of S_G is $1/2$, and each edge residual has coefficient sum $3/4$ with $|r_{ij}(X)| \le 1$ on the elliptope. The gradient of f_ρ^G therefore has coefficient sum at most $1/2 + 3\rho/2$, and its linear range in the theorem is at most $R = 1 + 3\rho$.

For the curvature, the tangent probe of that theorem has

$$\mathbb{E}H_{ij}^2 = \frac{2}{k}(1 - X_{ij}^2)(1 - X_{ij}(X_*)_{ij}) \le \frac{4}{k},$$

which follows by expanding H_{ij} into its two Gaussian linear forms, each of variance $(1 - X_{ij}^2)/k$ and with covariance $-X_{ij}(X_*)_{ij}(1 - X_{ij}^2)/k$. The L^2 triangle inequality applied to the coefficient sum $3/4$ then gives

$$\mathbb{E}(\mathrm{d}r_{ij}[H])^2 \le \frac{4}{k}\left(\frac{3}{4}\right)^2 = \frac{9}{4k},$$

and since

$$\nabla^2 f_\rho^G(X)[H,H] = -\frac{2\rho}{m}\sum_{ij\in E}(\mathrm{d}r_{ij}[H])^2,$$

the curvature constant is $C \le 9\rho/2$. Theorem 4.3(b) applied with a penalized maximizer as comparator gives the first bound in (147), with $R + C/2 \le 1 + 3\rho + 9\rho/4 = 1 + 21\rho/4$. Because feasible points have zero penalty, $c_\rho^G \ge \theta_* \ge 0$, and combining this with $F_\rho^G \le S_G \le 1$ and $\rho D_G = S_G - F_\rho^G$ gives the remaining bounds and, through Lemma L.5, the upper certificate. □

The proposition gives an upper endpoint for $\vartheta(G)/n$ from any factor with an objective Hessian bounded above by ηI, and a lower endpoint requires a feasible point of $\mathcal{T}_G$, which the current factor need not provide. Such a point is constructed below by an explicit repair whose cost is controlled by the residual, so that the two endpoints together form a finite interval.

Proposition L.7 (Explicit repair and a finite theta interval)**.** *Let $\mu = 1/[n(n+1)]$ and $\rho > 0$. For every $X \in \mathcal{E}_{n+1}$, one can construct $Z \in \mathcal{T}_G$ satisfying*

$$f_\rho^G(X) - S_G(Z) \le \frac{8m}{\mu^2 \rho}. \tag{148}$$

If $X = VV^\top$ also satisfies Proposition L.6, then

$$\begin{aligned} S_G(Z) \le \frac{\vartheta(G)}{n} \le F_\rho^G(V) + \Delta_G, \\ F_\rho^G(V) + \Delta_G - S_G(Z) \le \Delta_G + \frac{8m}{\mu^2 \rho}. \end{aligned} \tag{149}$$

Proof. The repair subtracts the edge residuals from X, which restores the edge equations but may destroy positive semidefiniteness. That matrix is then mixed with a positive definite feasible matrix W whose smallest eigenvalue μ controls how much mixing is needed. We construct W by mixing the independent-set Boolean lifts used in Lemma L.5, with weight $1/2$ on the empty set and $1/(2n)$ on each singleton, which gives

$$W_{ii} = 1, \qquad W_{0i} = 1/n - 1, \qquad W_{ij} = 1 - 2/n \quad (i \ne j,\ i,j \ge 1), \qquad S_G(W) = 1/(2n),$$

with vanishing residual on every edge. On the vertex subspace orthogonal to the all-ones vector its eigenvalue is $2/n$, and on the orthogonal two-dimensional complement it acts by the matrix

$$\begin{pmatrix} 1 & \sqrt{n}(1/n-1) \\ \sqrt{n}(1/n-1) & n-2+2/n \end{pmatrix},$$

which has positive trace at most $n+1$ and determinant $1/n$. The smaller eigenvalue of a positive definite 2×2 matrix is at least its determinant divided by its trace, so $W \succeq \mu I$.

Now define a symmetric correction matrix E by $\mathsf{E}_{ij} = \mathsf{E}_{ji} = 4r_{ij}(X)$ for $ij \in E$, with all other entries zero, including the anchor row and the diagonal. The matrix $Y = X - \mathsf{E}$ has diagonal one, satisfies every edge equation, and obeys $S_G(Y) = S_G(X)$ because the score reads only the anchor row. Put $d = \|\mathsf{E}\|_F = \sqrt{32 m D_G(X)}$; since $X \succeq 0$ and $\|\mathsf{E}\|_{\mathrm{op}} \le d$, we have $Y \succeq -dI$, so that

$$t = d/\mu, \qquad Z = \frac{Y + tW}{1+t}$$

is positive semidefinite while retaining diagonal one and all edge equations, and hence $Z \in \mathcal{T}_G$. Because the score is affine,

$$S_G(X) - S_G(Z) = \frac{t}{1+t}\big(S_G(X) - S_G(W)\big) \le t,$$

and substituting the value of t gives

$$f_\rho^G(X) - S_G(Z) \le \frac{\sqrt{32m}}{\mu}\sqrt{D_G(X)} - \rho D_G(X) \le \frac{8m}{\mu^2 \rho},$$

where the last step completes the square in $\sqrt{D_G(X)}$; this is (148). Feasibility of Z gives $S_G(Z) \le \vartheta(G)/n$, Proposition L.6 supplies the upper endpoint, and adding Δ_G to (148) proves the interval-width bound. □

To see the interval width as a function of the target accuracy, let $B_G = 8m/\mu^2$ and fix $0 < \varepsilon < 1$; the choices

$$\rho = 2B_G/\varepsilon, \qquad \eta \le \frac{\varepsilon}{2(n+1)}, \qquad k - 1 \ge \frac{4(1 + 21\rho/4)}{\varepsilon}$$

give an interval of width at most ε whenever the Hessian condition has been verified. Since $\mu^{-2} = n^2(n+1)^2$, these choices make the example's search rank $k - 1 = \Theta(mn^4/\varepsilon^2)$, so the constants are instance-dependent and large, and the repaired matrix may have full rank. The certificate bounds the theta relaxation and supplies no iteration bound for reaching the Hessian condition, while extracting an independent set requires a separate procedure; low-rank theta optimization itself is classical (Burer & Monteiro, 2003, Section 4.1).

L.8 Continuous optimization examples

The constants R and C in Theorem 4.3(b) also cover convex losses on correlations, which have no consistency penalty but a nonlinear objective. For normalized nonnegative weights w_{ij}, consider the pairwise loss $h(X) = \sum_{i<j} w_{ij}\ell_{ij}(X_{ij})$ with convex C^2 losses obeying $|\ell'_{ij}| \le L_1$ and $0 \le \ell''_{ij} \le L_2$ on $[-1,1]$, and apply the landscape bound to $f = -h$. The gradient coefficient sum is at most L_1, giving $R \le 2L_1$, and the entry bound (137) gives $-k\mathbb{E}\nabla^2 f[H,H] \le 4L_2$, so the entry in Table 4 holds whenever $\mathrm{Hess}(h \circ \mathrm{Gram}) \succeq -\eta I$. Squared fitting to supplied correlations and logistic pairwise losses are examples, and in each case the guarantee concerns the observed average loss.

For a contrastive loss, fix nonempty comparison sets J_i and probability vectors q_i supported on them, and define the temperature-τ objective

$$h_\tau(X) = \frac{1}{N}\sum_i \left[\log \sum_{j\in J_i} e^{X_{ij}/\tau} - \frac{1}{\tau}\sum_{j\in J_i} q_{ij}X_{ij}\right].$$

Writing p_i for the softmax probabilities of row i, the directed differential coefficients are $(p_{ij} - q_{ij})/(N\tau)$, whose absolute sum is at most $2/\tau$, so that $R \le 4/\tau$. The Hessian is a sum of softmax variances,

$$\nabla^2 h_\tau(X)[H,H] = \frac{1}{N\tau^2}\sum_i \mathrm{Var}_{j\sim p_i}(H_{ij}),$$

and bounding each variance by the second moment gives $C \le 4/\tau^2$, which is the contrastive entry of the table; the resulting guarantee compares in-sample loss with the convex Gram optimum.

L.9 A penalized semidefinite relaxation of K-means

The clustering relaxation is naturally posed on a trace-constrained set rather than on the elliptope, and this subsection shows that a doubled set of unit rows nevertheless represents it, at the price of a feasibility penalty. Here K is the number of clusters and k remains the search rank of the Gram factor. Let $x_1,\dots,x_n$ be data points with $n \ge 2$ and $1 \le K \le n$, and rescale the squared distances so that $d_{ij} = \|x_i - x_j\|^2 \le 1$, with $d_{ii} = 0$. The Peng–Wei semidefinite relaxation is

$$h_*^{\mathrm{PW}} = \min_{Z\succeq 0,\ Z\mathbf{1}=\mathbf{1},\ \mathrm{tr}\, Z = K,\ Z\ge 0} \frac{1}{2n}\sum_{i,j=1}^n d_{ij}Z_{ij}, \tag{150}$$

where $Z \ge 0$ is entrywise (Peng & Wei, 2007). For a partition into nonempty clusters, setting Z_{ij} to the reciprocal of the cluster size when i,j share a cluster and to zero otherwise gives a feasible matrix whose objective equals the average squared distance to the cluster means.

An elliptope lift. To represent the relaxation with unit rows, use $2n$ rows, two per data point, and read off the averaged block

$$Z_{ij}(X) = \frac{X_{i,j} + X_{i,n+j} + X_{n+i,j} + X_{n+i,n+j}}{4}, \qquad h(X) = \frac{1}{2n}\sum_{i,j} d_{ij}Z_{ij}(X) \tag{151}$$

for $X \in \mathcal{E}_{2n}$. If the unit rows representing X are $v_1,\dots,v_{2n}$, then $Z(X)$ is the Gram matrix of the midpoints $(v_i + v_{n+i})/2$, so $Z(X) \succeq 0$ and $Z_{ii}(X) \le 1$. Conversely, let $Z \succeq 0$ have diagonal at most one, choose vectors w_i with Gram matrix Z, and in additional orthogonal coordinates choose mutually orthogonal unit vectors e_i; the vectors

$$v_i = w_i + \sqrt{1 - \|w_i\|^2}\, e_i, \qquad v_{n+i} = w_i - \sqrt{1 - \|w_i\|^2}\, e_i$$

are then unit and reproduce Z through (151). Since every feasible matrix in (150) has diagonal at most one, by nonnegativity and its row sums, the lift represents the entire Peng–Wei feasible set.

The constraints become the affine residuals

$$a_i(X) = \frac{\sum_j Z_{ij}(X) - 1}{n+1}, \quad a_{\mathrm{tr}}(X) = \frac{\mathrm{tr}\, Z(X) - K}{2n}, \quad b_{ij}(X) = -Z_{ij}(X) \quad (i<j), \tag{152}$$

which encode the row-sum, trace, and sign constraints, and with $M = n + 1 + \binom{n}{2}$ residuals in all and $t_+ = \max\{t, 0\}$ we form the penalty and the penalized objective

$$\mathcal{D}(X) = \frac{1}{M}\left(\sum_i a_i(X)^2 + a_{\mathrm{tr}}(X)^2 + \sum_{i<j} b_{ij}(X)_+^4\right), \qquad h_\rho(X) = h(X) + \rho\mathcal{D}(X). \quad (153)$$

The penalty is convex and C^3, and its zero set maps onto the feasible set of (150). Let $h_{\rho,*} = \min_{X\in\mathcal{E}_{2n}} h_\rho(X)$ be the penalized optimum.

Proposition L.8 (A K-means relaxation certificate)**.** *Let $\rho > 0$, $k \geq 2$, $\eta \geq 0$, and $V \in \mathcal{M}_{2n,k}$. Suppose*

$$\mathrm{Hess}(h_\rho \circ \mathrm{Gram})(V) \succeq -\eta I, \qquad \delta = \frac{n + 32\rho}{k-1} + n\eta.$$

Then, with $X = VV^\top$,

$$0 \leq h_\rho(X) - h_{\rho,*} \leq \delta, \quad (154)$$

$$h(X) \leq h_*^{\mathrm{PW}} + \delta, \qquad \mathcal{D}(X) \leq \frac{n/2 + \delta}{\rho}, \quad (155)$$

$$h_\rho(X) - \delta \leq h_*^{\mathrm{PW}}. \quad (156)$$

The factor has $2nk$ stored coordinates and intrinsic dimension $2n(k-1)$.

Proof. We bound the linear range and the probe curvature of $f = -h_\rho$ by coefficient counting, apply the landscape bound, and then compare the penalized optimum with the Peng–Wei value through the lift. Written as a constant plus unordered off-diagonal coordinates of X, each affine residual has absolute coefficient sum at most one, the values being $(n-1/2)/(n+1)$ for a_i, $1/4$ for a_{tr}, and one for b_{ij}. Every residual also has magnitude at most one on $\mathcal{E}_{2n}$, because $|Z_{ij}| \leq 1$, $0 \leq \mathrm{tr}\, Z \leq n$, and $|\sum_j Z_{ij} - 1| \leq n$.

For $|t| \leq 1$, the first two derivatives of t^2 have magnitudes at most 2 and 2, while those of t_+^4 have magnitudes at most 4 and 12, so the off-diagonal coefficient sum of $\nabla\mathcal{D}$ is at most four. The coefficient sum of h is at most

$$L_h = \frac{1}{n}\sum_{i<j} d_{ij} \leq \frac{n-1}{2},$$

and the gradient range therefore satisfies $R(\nabla f) \leq 2L_h + 8\rho \leq n - 1 + 8\rho$. For the Gaussian tangent probe of Theorem 4.3(b), each affine residual a obeys $\mathbb{E}(da[H])^2 \leq 4/k$ by its coefficient sum, so that

$$-k\mathbb{E}\nabla^2 f(X)[H,H] = k\rho\mathbb{E}\nabla^2\mathcal{D}(X)[H,H] \leq 48\rho,$$

and the landscape bound gives (154), even with $n - 1 + 32\rho$ in place of $n + 32\rho$.

Since every feasible Peng–Wei matrix admits a lift with zero penalty, $h_{\rho,*} \leq h_*^{\mathrm{PW}}$, which proves the objective bound and (156). For the residual bound, note that $h(X) \geq -(n-1)/2$ on the elliptope. Any partition into K nonempty clusters has objective at most $(n-K)/(2n) \leq 1/2$, because its within-cluster distances are at most one, and hence

$$\rho\mathcal{D}(X) = h_\rho(X) - h(X) \leq h_*^{\mathrm{PW}} + \delta - h(X) \leq n/2 + \delta,$$

which is the second half of (155). □

The proposition certifies the penalized relaxation at a point that may still violate the Peng–Wei constraints. Equation (156) lower-bounds the optimal clustering cost, whereas an upper bound on that cost needs a feasible partition, and neither Lloyd's algorithm nor a rounding of the penalized output is analyzed here. Because the positive-part penalty is C^3 and semialgebraic rather than analytic, Proposition K.2 does not apply to it directly, and the certificate rests on the displayed Hessian condition alone.

M NEURAL PARAMETERIZATIONS AND THEIR HYPOTHESES

The unit-row bounds of Theorem 4.3(b) treat the factor rows as free variables, and Section 4.3 claims that they transfer to a neural encoder with a normalized linear head. This appendix supplies the proof of that transfer, examines when its full-rank feature hypothesis holds and how long training preserves it, and then records the two classical prediction-loss certificates listed in the last row of Table 4.

M.1 THE NORMALIZED-HEAD TRANSFER

For the two-layer Gram-training row of Table 4, the features are $\psi_i = (\sigma(w_j^\top x_i))_{j=1}^h$, the output head is $B \in \mathbb{R}^{k\times h}$, and there are $h(p+k)$ parameters for inputs in $\mathbb{R}^p$. At the inspected point we require $\operatorname{rank}\Psi = N$, hence $h \geq N$, together with $B\psi_i \neq 0$ for every row, and any Gram loss in the table then has the same landscape bound up to the head-conditioning tolerance of Corollary 4.5.

Proof of Corollary 4.5. The idea is to realize every tangent direction of the factor as a head perturbation through a left inverse of the feature matrix and to check that the normalized head outputs then follow the sphere geodesic to second order. For $U \in T_V\mathcal{M}_{N,k}$, let $Z_U = [\|z_i\|u_i]_i$ be the tangent direction rescaled row by row and put $\delta B = Z_U\Psi^+$; since full column rank gives $\Psi^+\Psi = I$, the perturbed head satisfies $\delta B\psi_i = \|z_i\|u_i \perp z_i$. Along the straight head perturbation $B + t\,\delta B$, row i of the normalized factor becomes

$$v_i(t) = \frac{v_i + tu_i}{\sqrt{1 + t^2\|u_i\|^2}},$$

whose first two derivatives at $t = 0$ are u_i and $-\|u_i\|^2 v_i$ and agree with those of the sphere geodesic, so that

$$\nabla_B^2[-F(V(B))][\delta B, \delta B] = -\operatorname{Hess} F(V)[U, U].$$

Since $\|\Psi^+\|_{\mathrm{op}} = 1/\sigma_N(\Psi)$, the perturbation obeys $\|\delta B\|_F \leq \max_i \|z_i\|\|U\|_F/\sigma_N(\Psi)$, and inserting this into the head Hessian bound gives the stated tolerance η. The first-derivative chain rule with the same arbitrary U shows that an exact stationary parameter point gives an exact stationary factor point, which completes the proof. □

Because only head directions are used, the backbone may be trained jointly, although the theorem remains a statement at the inspected parameter point. Its full-rank hypothesis fails whenever two feature columns coincide, and width alone does not repair this. An anonymous message-passing encoder with identical initial node features produces equal features on the cells of the coarsest equitable partition at every width. Distinct inputs do not suffice either, since a bias-free $\tanh$ backbone maps the inputs x and $-x$ to feature columns that are negatives of each other.

Some rows may also be optimized freely on their spheres while the others come from the encoder. The joint Hessian bound over head and free-row directions then gives the defect

$$\frac{N_{\mathrm{enc}}\gamma}{2}\left(\frac{\max_i \|z_i\|}{\sigma_{N_{\mathrm{enc}}}(\Psi)}\right)^2 + \frac{N_{\mathrm{free}}\gamma}{2},$$

in which only the encoder rows enter the full-rank feature condition. A bound on the separate diagonal Hessian blocks would be insufficient when mixed curvature is present.

M.2 A SUFFICIENT FINITE-HORIZON CONDITION

The transfer theorem needs $\sigma_N(\Psi) > 0$ at the inspected point, and one may ask whether training itself can be relied on to preserve it; the following conservation argument gives a sufficient condition over a finite horizon. Suppose that the features factor as $\Psi = B_1\Phi(\theta)$ through an inner linear layer $B_1 \in \mathbb{R}^{d\times d'}$ with $d \geq d' \geq N$ applied to backbone features $\Phi(\theta) \in \mathbb{R}^{d'\times N}$, and that a head B_2, playing the role of the head B of Corollary 4.5, acts on Ψ. Assume that the loss depends on B_2, B_1

only through the product B_2B_1 and that all outputs remain nonzero on the trajectory. Under joint Euclidean gradient flow the balance matrix

$$K = B_1B_1^\top - B_2^\top B_2$$

is constant (Du et al., 2018, Theorem 2.2), because writing $Q = \partial\ell/\partial(B_2B_1)$ gives $\nabla_{B_2}\ell = QB_1^\top$ and $\nabla_{B_1}\ell = B_2^\top Q$, so that the derivatives of $B_1B_1^\top$ and $B_2^\top B_2$ coincide. If $\lambda_{d'}(K(0)) = \lambda_0 > 0$, then the identity $B_1B_1^\top = K + B_2^\top B_2 \succeq K$ and Weyl monotonicity give $\sigma_{d'}(B_1(t)) \geq \sqrt{\lambda_0}$ for all t.

Under plain gradient descent the balance matrix is no longer conserved. For a step α satisfying the descent inequality $\ell_s - \ell_{s+1} \geq (\alpha/2)\|\nabla\ell_s\|^2$, direct expansion gives

$$K_{s+1} - K_s = \alpha^2\big(\nabla_{B_1}\ell_s\nabla_{B_1}\ell_s^\top - \nabla_{B_2}\ell_s^\top\nabla_{B_2}\ell_s\big),$$

so if $\ell_0 - \inf\ell \leq D_0$, then summing the descent inequality and using the same Weyl step yields $\sigma_{d'}(B_{1,s})^2 \geq \lambda_0 - 2\alpha D_0$. For an L_Φ-Lipschitz backbone and $T = \alpha s$, Cauchy–Schwarz and the same descent inequality bound the drift of the backbone features by

$$\sigma_N(\Phi_s) \geq \sigma_N(\Phi_0) - L_\Phi\sqrt{2TD_0},$$

and therefore, if $2\alpha D_0 < \lambda_0$ and $T \leq \sigma_N(\Phi_0)^2/(8L_\Phi^2 D_0)$,

$$\sigma_N(\Psi_s) \geq \tfrac{1}{2}\sqrt{\lambda_0 - 2\alpha D_0}\,\sigma_N(\Phi_0) > 0.$$

This is a conditional horizon guarantee for plain gradient descent. It keeps the feature matrix well conditioned but does not ensure that the required Hessian tolerance is reached within the horizon, and a perturbed method would need additional control of the perturbations in both recursions.

M.3 Unnormalized convex losses

The last row of Table 4 records two classical certificates for prediction losses that need no Gram factorization, and we include them because the same full-rank feature condition drives the first of them. For a scalar-output two-layer network, take $\psi_i = (\sigma(w_j^\top x_i))_{j=1}^h$ and the head $W = (a_1, \dots, a_h)$, so that with inputs in $\mathbb{R}^p$ the parameter count is $h(p+1)$; biases may be absorbed by augmenting the inputs. The full-rank condition below requires $h \geq N$ and $\operatorname{rank}\Psi = N$ at the inspected parameter point.

Let $Z = W\Psi$ be the prediction vector and let $h(Z)$ be differentiable and convex. If Ψ has full column rank, every output perturbation A is realized by the head perturbation $\delta W = A\Psi^+$, so that

$$\|\nabla_Z h(Z)\|_F \leq \frac{\|\nabla_W(h\circ Z)\|_F}{\sigma_N(\Psi)},$$

and a stationary head is therefore a global minimizer in prediction space. For a minimizer Z_*, convexity gives the quantitative version

$$h(Z) - h(Z_*) \leq \frac{\|\nabla_W(h\circ Z)\|_F\|Z - Z_*\|_F}{\sigma_N(\Psi)},$$

which is the familiar full-rank wide-layer mechanism of Nguyen & Hein (2017).

M.4 A classical certificate for regularized two-layer networks

The second classical certificate replaces the full-rank condition by positive homogeneity and a matched regularizer. Let $\phi(w) = (\sigma(w^\top x_i))_{i=1}^N$ be the response of one hidden unit on the training set, where σ is continuous and positively homogeneous of degree one, such as ReLU, and for a differentiable convex loss $h\colon \mathbb{R}^N \to \mathbb{R}$ and a regularization weight $\lambda > 0$ consider

$$J(a, w) = h\left(\sum_{j=1}^m a_j\phi(w_j)\right) + \frac{\lambda}{2}\sum_{j=1}^m(a_j^2 + \|w_j\|^2).$$

The proposition below specializes the homogeneous-network certificate of Haeffele & Vidal (2017) to this loss.

Proposition M.1 (Unused-neuron certificate)**.** *If a local minimum of J contains a neuron with $a_j = 0, w_j = 0$, then its value is globally optimal among networks of every finite width.*

Proof. Local minimality is used twice, once through the unused neuron to bound the dual variable and once through rescaling to relate the loss gradient to the regularizer, after which convexity compares with any competitor. Put $z = \sum_j a_j \phi(w_j)$ and $g = \nabla h(z)$. For a unit vector u, perturbing the unused neuron to $(a, w) = (t, tu)$ and to $(a, w) = (-t, tu)$ with $t > 0$ changes the output by $\pm t^2 \phi(u)$ and the regularizer by λt^2, so local minimality and differentiability of h give $|\langle g, \phi(u)\rangle| \le \lambda$ for every unit u. For any existing neuron, jointly rescaling a_j, w_j by $\sqrt{1+s}$ for small positive or negative s must have vanishing first derivative, which yields

$$\langle g, a_j \phi(w_j)\rangle + \frac{\lambda}{2}(a_j^2 + \|w_j\|^2) = 0,$$

and summing over neurons gives $\langle g, z\rangle + R = 0$, where R is the current regularizer. For any finite-width competitor $(\widetilde{a}, \widetilde{w})$ with output $\widetilde{z}$ and regularizer $\widetilde{R}$, positive homogeneity and the dual bound imply

$$\langle g, \widetilde{z}\rangle \ge -\lambda \sum_j |\widetilde{a}_j| \|\widetilde{w}_j\| \ge -\frac{\lambda}{2} \sum_j (\widetilde{a}_j^2 + \|\widetilde{w}_j\|^2) = -\widetilde{R},$$

and convexity of h then gives $h(\widetilde{z}) + \widetilde{R} \ge h(z) - \langle g, z\rangle = h(z) + R$, so the local minimum is globally optimal. □

The certificate depends on the unused neuron, on positive homogeneity, and on the matched regularizer, and it is separate from the C^2 Gram theorem, since its ReLU form need not have a classical Hessian at the unused neuron.