ACCELERATION AND ITS LIMITS IN GAUSSIAN VARIATIONAL INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

We study whether the classical momentum-based acceleration that improves the condition-number dependence of gradient descent from κ to $\sqrt{\kappa}$ for strongly convex, smooth objectives extends to Gaussian variational inference (GVI) for strongly log-concave, log-smooth target distributions. With exact Gaussian expectations, a convex composite formulation of GVI yields the classical accelerated $\sqrt{\kappa}$ convergence rate. Under pointwise derivative access, however, we construct a family of targets for which any randomized algorithm requires $\kappa^{1-o(1)}$ queries to attain a fixed accuracy from a uniformly bounded initial objective gap. Thus curvature bounds alone cannot guarantee uniform accelerated total-query complexity. We then ask which additional structure permits acceleration and give positive examples: under a Lipschitz-Hessian assumption, variance-reduced stochastic methods yield accelerated guarantees when the Hessian variation is sufficiently small. Finally, experiments on real datasets show substantial empirical speedups from acceleration across a broad range of methods.

1 INTRODUCTION

Variational inference (VI) approaches probabilistic inference through optimization: it fits a tractable probability law to a target distribution, providing an approximation that is widely used in Bayesian inference (Blei et al., 2017). This viewpoint makes momentum-based acceleration in optimization a natural source of faster inference algorithms. For an α-strongly convex, β-smooth objective, momentum improves gradient descent's complexity from $O(\kappa \log(1/\varepsilon))$ to $O(\sqrt{\kappa}\log(1/\varepsilon))$, where $\kappa = \beta/\alpha$ (Nesterov, 2004). With stochastic gradients, however, acceleration must also account for a variance-dependent cost, which can dominate the gains from momentum.

The parallel question has driven substantial progress in Markov chain Monte Carlo (MCMC), the other pillar of probabilistic inference alongside VI. Its development has also benefited from an optimization viewpoint: the law of overdamped Langevin diffusion is a Wasserstein gradient flow of KL divergence (Jordan et al., 1998), while underdamped Langevin connects sampling with momentum methods (Cheng et al., 2018; Ma et al., 2021); see Chewi (2026) for an overview. Acceleration in sampling is delicate and remains an active area of research. Recent developments begin to achieve sublinear condition-number dependence in algorithm complexity for strongly log-concave targets (Altschuler et al., 2026; Chen et al., 2026a; Lu & Luo, 2026), while a non-acceleration result holds for standard splitting schemes (Bou-Rabee, 2026).

In VI, momentum variants of black-box methods have been related to Langevin dynamics (Hoffman & Ma, 2020), and accelerated gradient methods have been analyzed for mean-field VI with exact expectations (Jiang et al., 2024). This paper takes a first step toward the theoretical foundations of momentum-based stochastic acceleration in VI, focusing on full-covariance Gaussian variational inference (GVI). For strongly log-concave, log-smooth targets, GVI admits a convex composite formulation whose expected potential inherits the target's curvature bounds (Challis & Barber, 2013; Domke, 2020). Exact Gaussian integration therefore makes classical acceleration available. With pointwise target access, however, the required expectations must be estimated, bringing the variance-dependent cost into play. For unbiased stochastic gradients with variance bounded by σ^2, the classical accelerated complexity is

$$O\left(\sqrt{\kappa}\log\frac{2\Delta_0}{\varepsilon} + \frac{\sigma^2}{\alpha\varepsilon}\right) \tag{1}$$

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.

(Ghadimi & Lan, 2013), where Δ_0 bounds the initial objective gap. In GVI, the variance bound depends on the target, the variational parameters, and the estimator, and must hold along the iterates. Establishing acceleration in the total number of target queries requires accounting for this dependence.

Our contributions are as follows:

- **An accelerated benchmark.** With exact Gaussian expectations, the convex composite formulation of GVI gives the classical $O(\sqrt{\kappa}\log(\Delta_0/\varepsilon))$ rate in both Cholesky and symmetric-root coordinates (Theorem 2).
- **A query lower bound.** We construct a family of strongly log-concave, log-smooth targets in dimension $\Theta((\log\kappa)^2)$, with uniformly bounded initial gaps, for which any randomized algorithm using derivatives of fixed finite order requires $\kappa^{1-o(1)}$ queries to attain a fixed expected accuracy uniformly over the family (Theorem 3). The lower bound therefore extends beyond any particular estimator or momentum scheme.
- **Acceleration with controlled Hessian variation.** Under a Lipschitz-Hessian assumption, we combine control variates and score corrections, whose variance is governed by the Hessian variation, with accelerated proximal and projected methods in gradient-only and Hessian variants. All retain the accelerated $O(\sqrt{\kappa}\log(2\Delta_0/\varepsilon))$ iteration count, and their total query complexity remains accelerated when the Hessian variation is small relative to the target accuracy (Section 4).

Numerical experiments illustrate the predicted condition-number scaling on Gaussian targets and the practical benefits of acceleration on graph-probit posteriors. Compared with recent progress on accelerated sampling, our results suggest that understanding stochastic acceleration in VI requires structure beyond log-concavity, together with efficient variance reduction that exploits it.

1.1 Related Work

In statistical inference, means and covariances are often the most used statistics, making Gaussian approximations and GVI useful in their own right. Gaussian mixtures, a natural extension, remain competitive with more elaborate variational samplers (Blessing et al., 2024). GVI provides a tractable starting point for understanding algorithmic behavior.

GVI admits a convex formulation in mean and covariance-factor parameters for log-concave targets (Challis & Barber, 2013; Domke, 2020), and its expected-potential term inherits smoothness from the potential V. These properties underpin convergence analyses of projected and proximal stochastic gradient methods (Domke et al., 2023; Kim et al., 2023).

Geometry has played an important role in the study of GVI. Under strong log-concavity and log-smoothness, geodesic strong convexity of the GVI objective in Bures–Wasserstein geometry yields algorithms with convergence guarantees that are linear or quadratic in the condition number (Lambert et al., 2022; Diao et al., 2023; Kim et al., 2026). Fisher–Rao geometry underlies natural gradient methods (Amari, 1998). For GVI, natural gradient algorithms and their convergence have been studied in both deterministic and stochastic settings (Chen et al., 2026b; Wu & Gardner, 2024; Kumar et al., 2025; Sun et al., 2025; Guilmeau et al., 2026). Momentum acceleration on manifolds requires controlling geometric distortion when comparing quantities in different tangent spaces (Ahn & Sra, 2020). Silver step-size schedules offer another approach to acceleration in Riemannian optimization (Park et al., 2025), with existing guarantees based on exact gradients. We focus on the flat geometry in the mean and covariance-factor coordinates and study momentum-based acceleration.

For stochastic VI, quadratic control variates (Miller et al., 2017; Geffner & Domke, 2020) and sticking-the-landing (STL) corrections (Roeder et al., 2017) reduce reparameterization gradient variance. Recent work has established convergence guarantees for variance-reduced GVI methods (Kim et al., 2024; Luu et al., 2025). We derive variance bounds in terms of Hessian variation and combine them with accelerated stochastic optimization (Ghadimi & Lan, 2013), obtaining accelerated total-query guarantees when Hessian variation is sufficiently small.

Existing theoretical work on GVI has largely focused on convergence guarantees. To the best of our knowledge, there are no prior impossibility results for GVI under pointwise derivative access to strongly log-concave and log-smooth targets. Query-complexity lower bounds have also appeared

in the literature on log-concave sampling (Chewi et al., 2022; 2024) where the accuracy is measured in total variation, as opposed to excess reverse KL in our work.

1.2 Problem Setup

Let $\pi_V(x) = Z_V^{-1} e^{-V(x)}$ with $Z_V := \int_{\mathbb{R}^d} e^{-V(x)} dx < \infty$ be a target density on $\mathbb{R}^d$, where $V \in C^2(\mathbb{R}^d)$, and let $\mathcal{G}_d$ denote the family of Gaussian distributions on $\mathbb{R}^d$:

$$\mathcal{G}_d := \left\{ \mathcal{N}(m, \Sigma) : m \in \mathbb{R}^d,\, \Sigma \in \mathbb{S}^d_{++} \right\}.$$

For $q \in \mathcal{G}_d$, define

$$F_V(q) := \mathbb{E}_{X\sim q}[V(X)] - H(q), \qquad H(q) := -\mathbb{E}_{X\sim q}[\log q(X)].$$

where H is the entropy of q. The Gaussian variational inference (GVI) problem is to minimize F_V over $\mathcal{G}_d$. We suppress the subscript and write $F := F_V$ and $F^\star := \inf_{q\in\mathcal{G}_d} F(q)$. Because $F(q) = \mathrm{KL}(q \,\|\, \pi_V) - \log Z_V$, the quantity $F(q) - F^\star$ is exactly the excess reverse KL divergence relative to the best Gaussian approximation to π_V. We make the following assumptions on V:

Assumption 1. There are constants $0 < \alpha \le \beta < \infty$ such that

$$\alpha I \preceq \nabla^2 V(x) \preceq \beta I \qquad \text{for every } x \in \mathbb{R}^d.$$

This is the standard setting for non-asymptotic guarantees in log-concave sampling and in GVI (Chewi, 2026; Lambert et al., 2022; Diao et al., 2023; Domke et al., 2023). Throughout the paper, we write $\kappa := \beta/\alpha$. Under Assumption 1, F has a unique minimizer $q_\star = \mathcal{N}(m_\star, \Sigma_\star)$ over $\mathcal{G}_d$, and

$$\mathbb{E}_{q_\star}[\nabla V(X)] = 0, \qquad \mathbb{E}_{q_\star}[\nabla^2 V(X)] = \Sigma_\star^{-1}, \qquad \beta^{-1} I \preceq \Sigma_\star \preceq \alpha^{-1} I.$$

However, in Section 3, we show that these assumptions alone are insufficient to obtain full $\sqrt{\kappa}$ acceleration. The stochastic acceleration results later in the paper additionally use the following regularity condition:

Assumption 2. The potential V has a Lipschitz continuous Hessian. That is,

$$L_H := \sup_{x,y\in\mathbb{R}^d,\, x\neq y} \frac{\|\nabla^2 V(x) - \nabla^2 V(y)\|_{\mathrm{op}}}{\|x - y\|} < \infty.$$

The constant L_H quantifies the nonquadraticity of V. In particular, if $L_H = 0$, then V is quadratic and π_V is Gaussian.

We represent covariance Σ either by its Cholesky factor $\Sigma = LL^\top$, where $L \in \mathbb{T}^d_{++}$ is lower triangular with strictly positive diagonal, or by its symmetric square root $\Sigma = C^2$, where $C \in \mathbb{S}^d_{++}$ is positive definite. We write $\mathcal{A} = \mathbb{T}^d$ or $\mathbb{S}^d$, respectively, and let $\mathcal{A}_{++}$ denote the corresponding positive domain. We write $A \in \mathcal{A}_{++}$ for either factor.

For $A \in \mathcal{A}_{++}$, let $X = m + A\xi$ with $\xi \sim \mathcal{N}(0, I_d)$. Up to an additive constant independent of (m, A),

$$F(m, A) = \mathbb{E}_\xi[V(m + A\xi)] - \log\det A.$$

We use $x = (m, A)$ when the factor coordinate is clear from context and equip the product space with the norm $\|(u, U)\|^2 := \|u\|^2 + \|U\|_F^2$.

The gradients with respect to C are taken in $\mathbb{S}^d$ under the Frobenius inner product. For a differentiable function Φ on $\mathbb{S}^d_{++}$, $\nabla_C \Phi(C)$ denotes the unique symmetric matrix satisfying $D\Phi(C)[U] = \langle \nabla_C \Phi(C), U\rangle_F$ for every $U \in \mathbb{S}^d$.

Under these assumptions, a key observation is that the curvature of V is inherited by $\mathbb{E}_\xi[V(m+A\xi)]$, which is formalized by the following lemma (Challis & Barber, 2013; Domke, 2020).

Lemma 1. *Under Assumption 1, the energy function*

$$E(m, A) := \mathbb{E}_\xi[V(m + A\xi)]$$

is twice continuously differentiable on $\mathbb{R}^d \times \mathcal{A}$ and, for every $\theta = (m, A)$ and every direction (u, U),

$$\alpha\|(u, U)\|^2 \le \nabla^2 E(\theta)[(u, U), (u, U)] \le \beta\|(u, U)\|^2.$$

Consequently, E is α-strongly convex and β-smooth.

The proof is in Appendix A, where we use the isotropy identity: $\mathbb{E}\|u + U\xi\|^2 = \|u\|^2 + \|U\|_F^2$.

Access model and cost. Unless stated otherwise, algorithms access V through pointwise derivative evaluations. We count target-gradient evaluations and Hessian–vector products (HVPs) separately as $N_{\nabla V}$ and N_{HVP}, forming $\nabla^2 V(X)A$ costs d HVPs. An output q attains accuracy ε if $\mathbb{E}[F(q) - F^\star] \le \varepsilon$, and T_ε denotes the number of updates after which this holds. Section 2 alone uses exact Gaussian integration as an idealized benchmark.

2 Acceleration with Deterministic Gradients

Suppose the Gaussian expectations can be evaluated exactly: one exact-expectation call at (m, A) returns

$$\nabla E(m, A) = \big(\mathbb{E}[\nabla V(m + A\xi)],\, P_{\mathcal{A}}\mathbb{E}[\nabla V(m + A\xi)\xi^\top]\big), \qquad P_{\mathbb{T}^d} = \mathrm{low}, \quad P_{\mathbb{S}^d} = \mathrm{sym}$$

where $\mathrm{sym}(\cdot)$ and $\mathrm{low}(\cdot)$ denote the symmetric and lower-triangular projections, respectively. GVI is then a deterministic convex composite problem with the extended objective $F = E + h$, where $h(A) = -\log\det A$ on $\mathcal{A}_{++}$ and $h(A) = +\infty$ otherwise.

By Lemma 1, E is α-strongly convex and β-smooth on the entire ambient space, while h is proper, closed, and convex. Moreover, for any stepsize $\eta > 0$, the proximal map $\mathrm{prox}_{\eta h}(\cdot)$ is explicit: for $\mathcal{A} = \mathbb{T}^d$, it applies $y \mapsto \phi_\eta(y) := (y + \sqrt{y^2 + 4\eta})/2$ to the diagonal entries and the identity map to the strictly lower-triangular entries, while for $\mathcal{A} = \mathbb{S}^d$, it applies ϕ_η to the eigenvalues. This formula also appears in Domke (2020). The accelerated proximal gradient (APG) method therefore applies directly:

$$x_{k+1} = \mathrm{prox}_{\eta h}\big(y_k - \eta\nabla E(y_k)\big), \quad y_{k+1} = x_{k+1} + \rho(x_{k+1} - x_k),$$

with $x_k, y_k \in \mathbb{R}^d \times \mathcal{A}$, stepsize $\eta = 1/\beta$, and momentum $\rho = (1 - \sqrt{\alpha/\beta})/(1 + \sqrt{\alpha/\beta})$. The standard APG guarantee yields the following bound.

Theorem 2. *Let Assumption 1 hold, and initialize the iteration above at $x_0 = (m_0, A_0) \in \mathbb{R}^d \times \mathcal{A}_{++}$. For $x_k = (m_k, A_k)$, let $q_k := \mathcal{N}(m_k, A_k A_k^\top)$, let $x_\star$ be the minimizer, and define $\Delta_0 := F(q_0) - F^\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2$. For either $\mathcal{A} \in \{\mathbb{T}^d, \mathbb{S}^d\}$ and every $0 < \varepsilon \le \Delta_0$, the iterates satisfy*

$$F(q_{T_\varepsilon}) - F^\star \le \varepsilon, \qquad T_\varepsilon = O\left(\sqrt{\kappa}\log\frac{\Delta_0}{\varepsilon}\right).$$

The full contraction statement, precise parameter choices, and proof are given in Appendix A.

3 Impossibility of Uniform Stochastic Acceleration for GVI

Section 2 attains the accelerated $O(\sqrt{\kappa})$ dependence with exact Gaussian expectations. With pointwise access, where V and its derivatives are evaluated only at individual query points, these expectations must instead be estimated, for example by Monte Carlo. Then, a direct variance bound loses the acceleration: since ∇V is β-Lipschitz, the mean component of the standard reparameterization estimator has variance of order $\beta^2\,\mathrm{tr}(\Sigma)$, so the resulting bound on the variance term in equation 1 scales as κ^2. Better estimators may reduce this cost, but we show that under Assumption 1 alone, no algorithm with pointwise access can avoid a nearly linear dependence on κ in the total number of target evaluations.

For the lower bound, we allow the algorithm to evaluate V and its derivatives of any fixed finite order at each query point. We count each evaluation as one query, including evaluations within a Monte Carlo batch. Each point may be chosen based on previous observations and the algorithm's internal randomness. This includes any Monte Carlo method built from gradient or HVP evaluations. By rescaling the variables $y = \sqrt{\alpha}\,x$, we may normalize $\alpha = 1$ and consider the class

$$\mathcal{V}_d(\kappa) := \left\{V \in C^\infty(\mathbb{R}^d) : I_d \preceq \nabla^2 V(x) \preceq \kappa I_d \text{ for all } x \in \mathbb{R}^d\right\}.$$

All asymptotic notation in this section is as $\kappa \to \infty$.

Theorem 3. *There exist constants $\varepsilon_0, C_0, C > 0$ and $\kappa_0 > 1$ such that, for every $\kappa \ge \kappa_0$, there exist a dimension $d = \Theta((\log\kappa)^2)$ and a family $\{V_v : v \in \mathbb{S}^{d-1}\} \subseteq \mathcal{V}_d(\kappa)$ satisfying*

$$\sup_{v\in\mathbb{S}^{d-1}} \left(F_{V_v}(\mathcal{N}(0, I_d)) - F^\star_{V_v}\right) \le C_0.$$

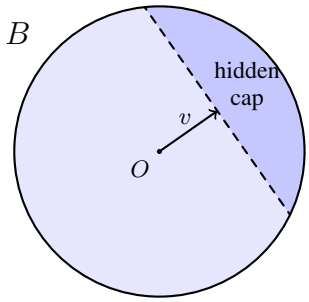


Figure 1: A hidden cap in direction v inside the ball B.

Furthermore, for every finite order $r \geq 1$, any randomized algorithm that makes at most N queries, evaluating V_v and its derivatives up to order r at each query point, and returns $q \in \mathcal{G}_d$ satisfying

$$\sup_{v \in \mathbb{S}^{d-1}} \mathbb{E}\left[F_{V_v}(q) - F^\star_{V_v}\right] \leq \varepsilon_0$$

must satisfy

$$N \geq \kappa \exp\left[-C(\log \kappa)^{3/4}\right].$$

Since $d = \Theta((\log \kappa)^2)$, any fixed polynomial in d is $\kappa^{o(1)}$. At the fixed accuracy ε_0, Theorem 3 therefore rules out every uniform bound of order $\widetilde{O}(\mathrm{poly}(d)\kappa^p)$ with $p < 1$, and in particular every accelerated bound of the form $O(\mathrm{poly}(d)\sqrt{\kappa}\,(\log(\Delta_0/\varepsilon) + 1/\varepsilon))$. The bounded initial gap ensures that the obstruction is not due to a poor initialization.

Proof sketch. Fix a ball B centered at the origin and an unknown direction $v \in \mathbb{S}^{d-1}$. The hyperplane $v^\top x = R$ cuts off a small cap of B in the direction v, as illustrated in Figure 1. We construct a reference potential V_B, which depends only on B, and for each v, a potential V_v whose values and all derivatives of any fixed finite order agree with those of V_B whenever the projection onto B stays away from the hidden cap. The cap parameters and the dimension d are chosen to achieve two properties. First, the standard Gaussian $\mathcal{N}(0, I_d)$ remains a constant objective gap away from the optimum for every potential V_v, uniformly over the hidden direction v. Second, for any fixed query point, V_v and V_B differ there only for a small fraction of directions v, of probability at most $C\kappa^{-1}\exp[C(\log \kappa)^{3/4}]$.

Now fix the algorithm's internal randomness. Its run against V_v coincides with its run against V_B until a query reaches the hidden region, and the reference queries do not depend on v. A union bound over at most N queries shows that the hidden region is reached with probability at most $CN\kappa^{-1}\exp[C(\log \kappa)^{3/4}]$. Otherwise, the output equals the reference output, which does not depend on v, and Lemma 14 (Appendix B) shows that any fixed Gaussian has objective gap at most $2\varepsilon_0$ for at most a $\kappa^{-\Omega(1)}$ fraction of directions. Hence, with $\mathbb{P}_{v,\mathrm{alg}}$ denoting probability over v and the algorithm's randomness,

$$\mathbb{P}_{v,\mathrm{alg}}\left(F_{V_v}(q) - F^\star_{V_v} \leq 2\varepsilon_0\right) \leq C\frac{N}{\kappa}\exp\left[C(\log \kappa)^{3/4}\right] + \kappa^{-\Omega(1)}.$$

If the worst-case expected gap is at most ε_0, Markov's inequality makes the left-hand side at least $1/2$. For $\kappa \geq \kappa_0$ the last term is at most $1/4$, and rearranging gives $N \geq \kappa\exp[-C(\log \kappa)^{3/4}]$ after enlarging C. Appendix B gives the complete proof.

Remark 1. The construction has large Hessian variation, so it does not contradict the accelerated guarantees of the next section, which require this variation to be small.

4 Recovering Stochastic Acceleration with Small Hessian Variation

Section 3 shows that Assumption 1 alone is insufficient for uniform stochastic acceleration. We now give positive examples under Assumption 2: accelerated dependence on κ can be recovered

when the Hessian variation is sufficiently small. We give two complementary constructions. The first uses a gradient-only finite-difference (FD) control variate in Cholesky coordinates, and the second uses score corrections in symmetric-root coordinates. For an initial point $x_0 = (m_0, A_0)$, let $q_0 := \mathcal{N}(m_0, A_0 A_0^\top)$ and let $x_\star$ denote the minimizer. Throughout this section, write

$$\Delta_0 := F(q_0) - F^\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2, \qquad \lambda_H := \frac{L_H}{\alpha^{3/2}}.$$

Note that after normalizing $\alpha = 1$, λ_H is simply the Hessian-Lipschitz constant.

4.1 Finite-difference control variates in Cholesky coordinates

For $\nu \neq 0$, define

$$D_\nu(m; e_j) := \frac{\nabla V(m + \nu e_j) - \nabla V(m)}{\nu}, \qquad D_\nu(m) := \big[D_\nu(m; e_1) \ \cdots \ D_\nu(m; e_d)\big].$$

The matrix $D_\nu(m)$ is a first-order approximation of $\nabla^2 V(m)$ and is exact whenever V is quadratic. For $\xi \sim \mathcal{N}(0, I_d)$, let

$$r_\nu(m, L; \xi) := \nabla V(m + L\xi) - \nabla V(m) - D_\nu(m) L\xi$$

and define

$$G^{\mathrm{FD}}(m, L; \xi) := \left(\nabla V(m) + r_\nu(m, L; \xi), \operatorname{low}\left(D_\nu(m)L + r_\nu(m, L; \xi)\xi^\top\right)\right).$$

For a batch of size b, G_b^{FD} denotes the average of b independent samples using a common $D_\nu(m)$. Thus, $d + 1$ gradients are shared across the batch, and one update uses $d + 1 + b$ first-order oracle calls. The FD estimator G^{FD} is unbiased for ∇E, and its variance admits a bound proportional to L_H^2 (Proposition 15).

To obtain a variance bound along the iterates, we impose the redundant restriction $\|L\|_F \le \sqrt{d/\alpha}$ and define

$$\bar{h}(L) := \begin{cases} -\log\det L, & L \in \mathbb{T}^d_{++} \text{ and } \|L\|_F \le \sqrt{d/\alpha}, \\ +\infty, & \text{otherwise.} \end{cases}$$

This restriction does not change the optimum, since $\|L_\star\|_F^2 = \operatorname{tr}(\Sigma_\star) \le d/\alpha$. We then apply APG to $E + \bar{h}$, using G_b^{FD} in place of the exact energy gradient. Proposition 16 (in Appendix C.1) shows that the resulting proximal step reduces to solving a univariate monotone equation.

Algorithm 1 Cholesky proximal acceleration

Require: $x_0 = (m_0, L_0)$ with $L_0 \in \operatorname{dom}\bar{h}$, batch size b, FD scale ν, horizon T

1: $y_0 \leftarrow x_0$
2: **for** $k = 0, \dots, T-1$ **do**
3: $\quad g_k \leftarrow G_b^{\mathrm{FD}}(y_k)$
4: $\quad (m_{k+1}, \widetilde{L}_{k+1}) \leftarrow y_k - \eta g_k$
5: $\quad x_{k+1} \leftarrow (m_{k+1}, \operatorname{prox}_{\eta\bar{h}}(\widetilde{L}_{k+1}))$, $y_{k+1} \leftarrow x_{k+1} + \rho(x_{k+1} - x_k)$
6: **end for**
7: **return** $q_T = \mathcal{N}(m_T, L_T L_T^\top)$

The fixed stepsize η and momentum ρ are specified in Theorem 18. The resulting accelerated convergence guarantee is as follows.

Theorem 4. *Suppose Assumptions 1 and 2 hold, $x_0 \in \operatorname{dom}\bar{h}$, $0 < |\nu| \le \alpha^{-1/2}$, and $0 < \varepsilon \le \Delta_0$. There is a fixed batch size b such that Algorithm 1 satisfies*

$$\mathbb{E}[F(q_{T_\varepsilon}) - F^\star] \le \varepsilon, \qquad T_\varepsilon = O\left(\sqrt{\kappa}\log\frac{2\Delta_0}{\varepsilon}\right), \qquad N_{\nabla V} = O\left(\left(d\sqrt{\kappa} + \frac{\lambda_H^2 d^3}{\varepsilon}\right)\log\frac{2\Delta_0}{\varepsilon}\right).$$

The full constants are given in Theorem 18. The same update also accepts the second-order estimator G^{paired}, which evaluates the gradient and the Hessian at independent Gaussian samples (Proposition 19 and Corollary 20). The finite-difference estimator can also be combined with a Euclidean projection, with the same asymptotic first-order complexity; see Appendix C.2.

4.2 SCORE CORRECTIONS IN SYMMETRIC-ROOT COORDINATES

The finite-difference method uses only first-order derivatives but requires $d+1$ additional calibration queries per update. We next use Gaussian score corrections to obtain relative-noise bounds without this additional cost. For $q = \mathcal{N}(m, C^2)$ with $C \in \mathbb{S}^d_{++}$ and $X := m + C\xi$, $\xi \sim \mathcal{N}(0, I_d)$, we have $\nabla_x \log q(X) = -C^{-1}\xi$, so the sticking-the-landing correction (Roeder et al., 2017) gives

$$G^{\mathrm{STL}}(m, C; \xi) := \left(\nabla V(X) - C^{-1}\xi, \mathrm{sym}\left((\nabla V(X) - C^{-1}\xi)\xi^\top\right)\right).$$

Using Stein's identity (Stein, 1981; Price, 1958)

$$\mathbb{E}[\nabla V(X)\xi^\top] = \mathbb{E}[\nabla^2 V(X) C],$$

we define the Price–STL estimator (Luu et al., 2025; Kim et al., 2026)

$$G^{\text{Price-STL}}(m, C; \xi) := \left(\nabla V(X) - C^{-1}\xi, \mathrm{sym}(\nabla^2 V(X) C) - C^{-1}\right).$$

Both estimators are unbiased for the full gradient $\nabla F(m, C)$. To control the inverse factor and bound the estimator moments, we restrict the iterates to

$$K_C := \left\{(m, C) : m \in \mathbb{R}^d,\ C \in \mathbb{S}^d,\ \frac{1}{2\sqrt{\beta}} I \preceq C \preceq \frac{2}{\sqrt{\alpha}} I\right\}.$$

The optimal symmetric factor lies in the interior of K_C. For a batch of size b, let G_b denote the average of b independent samples of either score estimator G^{STL} or $G^{\text{Price-STL}}$. On K_C, both estimators have second moments bounded by a multiple of $\|\nabla F\|^2$ plus an additive term proportional to L_H^2/b (Lemma 26). We use the following projected acceleration scheme with fixed parameters η, s (Theorem 29). See Appendix C.3 for details.

Algorithm 2 Projected acceleration with STL or Price–STL

Require: $x_0 \in K_C$, batch size b, estimator $G_b \in \{G_b^{\mathrm{STL}}, G_b^{\text{Price-STL}}\}$, horizon T

$v_0 \leftarrow x_0$, $y_0 \leftarrow x_0$
for $k = 0, \ldots, T-1$ **do**
  $g_k \leftarrow G_b(y_k)$
  $x_{k+1} \leftarrow P_{K_C}(y_k - \eta g_k)$
  $v_{k+1} \leftarrow P_{K_C}((1 - s^{-1})v_k + s^{-1}y_k - (\alpha s)^{-1} g_k)$
  $y_{k+1} \leftarrow (x_{k+1} + s^{-1} v_{k+1})/(1 + s^{-1})$
end for
return $q_T = \mathcal{N}(m_T, C_T^2)$

Both projected sequences remain in K_C, and each query point is their convex combination. Hence the inverse factor and both score estimators remain well-defined throughout the run.

Theorem 5. *Suppose Assumptions 1 and 2 hold, $x_0 \in K_C$, and $0 < \varepsilon \le \Delta_0$. For G^{STL}, there is a fixed batch size such that Algorithm 2 attains*

$$\mathbb{E}[F(q_{T_\varepsilon}) - F^\star] \le \varepsilon, \qquad T_\varepsilon = O\left(\sqrt{\kappa}\log\frac{2\Delta_0}{\varepsilon}\right), \qquad N_{\nabla V} = O\left(\left(d\sqrt{\kappa} + \frac{\lambda_H^2 d^3}{\varepsilon}\right)\log\frac{2\Delta_0}{\varepsilon}\right).$$

Table 1 summarizes the fixed batch sizes and oracle complexities. See Theorem 29 for the full second-order guarantees.

Interpretation. The two first-order methods use gradient evaluations differently: FD uses $d+1$ calibration gradients per update to reduce variance, while STL avoids calibration but requires a batch size proportional to d (Table 1).

When $\lambda_H^2 d^2 \lesssim \varepsilon\sqrt{\kappa}$, the accelerated term dominates every oracle count in Table 1. Second-order estimators then need only constant batches, and the first-order ones batches of size $O(d)$. This complements Section 3: curvature bounds alone do not permit uniform acceleration, while controlled Hessian variation provides a guarantee for acceleration.

We have not optimized the dimension dependence. Existing analyses of full-covariance GVI give iteration bounds with explicit factors linear in d (Domke et al., 2023; Kim et al., 2024; 2026). For Price–STL, a sharper variance analysis under additional structural assumptions also yields a gradient-query sampling term linear in d (Appendix C.4).

Method	Batch size	Gradient queries	HVP queries
FD + Alg. 1	$1 + \frac{\lambda_H^2 d^3}{\varepsilon\sqrt{\kappa}}$	$d\sqrt{\kappa} + \frac{\lambda_H^2 d^3}{\varepsilon}$	0
STL + Alg. 2	$d + \frac{\lambda_H^2 d^3}{\varepsilon\sqrt{\kappa}}$	$d\sqrt{\kappa} + \frac{\lambda_H^2 d^3}{\varepsilon}$	0
Paired + Alg. 1	$1 + \frac{\lambda_H^2 d^2}{\varepsilon\sqrt{\kappa}}$	$\sqrt{\kappa} + \frac{\lambda_H^2 d^2}{\varepsilon}$	$d\sqrt{\kappa} + \frac{\lambda_H^2 d^3}{\varepsilon}$
Price–STL + Alg. 2	$1 + \frac{\lambda_H^2 d^2}{\varepsilon\sqrt{\kappa}}$	$\sqrt{\kappa} + \frac{\lambda_H^2 d^2}{\varepsilon}$	$d\sqrt{\kappa} + \frac{\lambda_H^2 d^3}{\varepsilon}$

Table 1: Fixed-batch guarantees. All four methods (estimator, algorithm) attain expected objective gap at most ε in $T_\varepsilon = O\left(\sqrt{\kappa}\log(2\Delta_0/\varepsilon)\right)$ iterations. Entries give orders in $O(\cdot)$ notation; the two oracle-count columns omit the common factor $\log(2\Delta_0/\varepsilon)$.

5 Experiments

In this section, we present experimental results for our accelerated methods. Each method is labeled with the estimator, proximal/projected update (prox/proj), and accelerated/nonaccelerated variant (acc./nonacc.) that is used. We refer to Appendix D for full experimental details.

Gaussian targets. We compare Algorithm 2 using G_b^{STL} with its nonaccelerated projected-gradient counterpart, at $b = 64$. We use the standard normal distribution initialization $q_0 = \mathcal{N}(0, I_d)$, and vary the condition number κ from 10 to 3000. Figure 2 confirms the predicted scaling: the iteration count grows as $\sqrt{\kappa}$ with momentum and as κ without. Appendix D.1 also reports the samplewise-exact Paired benchmark and comparisons with all seventeen methods at $\kappa = 300$.

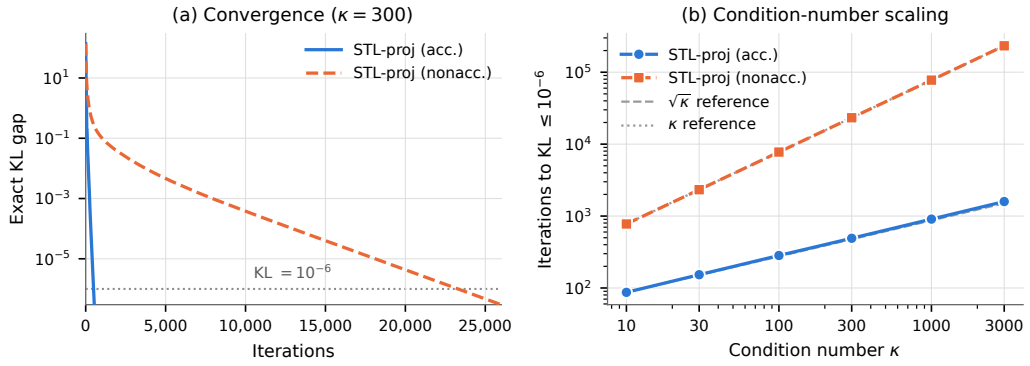


Figure 2: **Stochastic acceleration on Gaussian targets.** STL-proj with $b = 64$: (a) mean KL gap at $\kappa = 300$ with 20–80 percentile bands over five seeds; (b) updates to mean KL $\leq 10^{-6}$, with $\sqrt{\kappa}$ and κ reference lines.

Non-Gaussian targets. We construct graph-probit posteriors (Miller et al., 2020; Bertozzi et al., 2018) from five binary-classification datasets in OpenML (Vanschoren et al., 2014). Standardized predictors define a 10-nearest-neighbor graph with normalized Laplacian L_G. We use one latent score per observation and Gaussian prior precision $Q_\tau = I_n + \tau L_G$, with chosen hyperparameter $\tau = 49$. Graph-probit allows conditioning to worsen without increasing Hessian variation: increasing τ raises $\kappa = 2 + \tau\lambda_{\max}(L_G)$, while L_H remains independent of τ. Refer to Appendix D.2 for further details.

We compare seventeen first- and second-order methods, including five baselines (Domke et al., 2023; Diao et al., 2023; Luu et al., 2025; Kim et al., 2026), from $\mathcal{N}(0, I_d)$ for 600 updates with five seeds. FD and paired use proximal Algorithm 1 or projected Algorithm 4, and STL and Price-STL use Algorithm 2. The matching variants share the estimator, stepsize, and derivative cost per update.

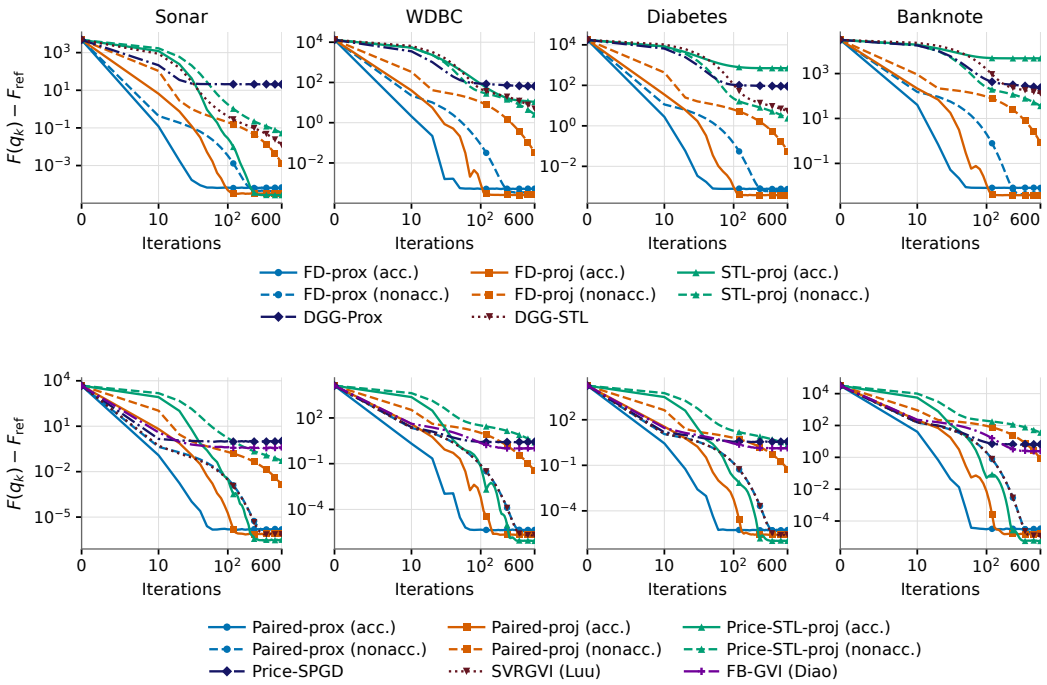


Figure 3: **Convergence on graph-probit posteriors.** The top row compares first-order methods and the bottom row compares second-order methods, all with batch size 64. Curves are five-seed mean objective gaps from $\mathcal{N}(0, I_d)$, with 20–80 percentile bands. Solid and dashed lines distinguish accelerated and nonaccelerated variants within each matched pair. Appendix D.5 reports comparisons across all five datasets.

Appendix D.5 also compares methods by gradient-query cost. Objective gaps use deterministic quadrature and a common high-accuracy numerical reference $F_{\text{ref}} \approx F^\star$ (Appendix D.4).

Results. At $b = 64$, every accelerated variant except STL-proj reaches 10^{-5} of the initial mean objective gap before its nonaccelerated counterpart on all five datasets (Table 4). Consistent with its dimension-dependent batch requirement, momentum can hurt STL-proj at small batches. At $b = 256$, it reaches the 1% threshold sooner and attains smaller final gaps than its nonaccelerated counterpart on all five datasets (Appendix D.5).

6 Conclusion

This work takes a first step toward the theoretical foundations of momentum-based stochastic acceleration in Gaussian variational inference. Exact Gaussian expectations permit classical $\sqrt{\kappa}$ acceleration, but, perhaps surprisingly, curvature bounds alone cannot ensure uniform acceleration under pointwise access. This contrasts with recent progress on accelerated sampling for strongly log-concave targets (Altschuler et al., 2026; Chen et al., 2026a; Lu & Luo, 2026). Our variance-reduced stochastic methods recover accelerated guarantees under sufficiently small Hessian variation, and experiments show empirical gains on real-data posteriors. Future work includes identifying which structures determine stochastic acceleration, and comparing systematically with geometry-aware methods such as natural gradient descent, which are subject to the same lower bound yet often perform well in practice, suggesting that they exploit different favorable structure.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting and revising the manuscript, implementing and debugging the experimental code, and writing and checking proofs. These tools also supported literature search, experiment design, and the analysis and presentation of experimental results. The research questions, the mathematical claims, and the main ideas were developed by the authors. All AI-assisted text, code, proofs, and experimental results were reviewed by the authors. We take responsibility for the final content of this work, including the text, claims, proofs, and artifacts produced with AI assistance.

REFERENCES

Kwangjun Ahn and Suvrit Sra. From Nesterov's estimate sequence to Riemannian acceleration. *Conference on Learning Theory*, 2020.

Jason M. Altschuler, Sinho Chewi, and Matthew S. Zhang. Shifted composition IV: Toward ballistic acceleration for log-concave sampling. *ACM Symposium on Theory of Computing*, 2026.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Andrea L. Bertozzi, Xiyang Luo, Andrew M. Stuart, and Konstantinos C. Zygalakis. Uncertainty quantification in graph-based classification of high dimensional data. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):568–595, 2018.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Denis Blessing, Xiaogang Jia, Johannes Esslinger, Francisco Vargas, and Gerhard Neumann. Beyond ELBOs: A large-scale evaluation of variational methods for sampling. *International Conference on Machine Learning*, 2024.

Nawaf Bou-Rabee. Provable non-acceleration of standard Strang splittings of kinetic Langevin dynamics. *arXiv:2608.25279*, 2026.

Edward Challis and David Barber. Gaussian Kullback–Leibler approximate inference. *Journal of Machine Learning Research*, 14(68):2239–2286, 2013.

Fan Chen, Sinho Chewi, Alexander Rakhlin, and Matthew S. Zhang. Exact simulation of diffusions and improved algorithms for log-concave sampling. *arXiv:2608.05022*, 2026a.

Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, and Andrew M. Stuart. Sampling via gradient flows in the space of probability measures. *Mathematics of Computation*, 2026b.

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. *Conference on Learning Theory*, 2018.

Sinho Chewi. Log-concave sampling, 2026. URL `https://chewisinho.github.io/main.pdf`. Book draft, version August 20, 2026.

Sinho Chewi, Patrik R. Gerber, Chen Lu, Thibaut Le Gouic, and Philippe Rigollet. The query complexity of sampling from strongly log-concave distributions in one dimension. *Conference on Learning Theory*, 2022.

Sinho Chewi, Jaume de Dios Pont, Jerry Li, Chen Lu, and Shyam Narayanan. Query lower bounds for log-concave sampling. *Journal of the ACM*, 71(4):29:1–29:42, 2024. doi: 10.1145/3673651.

Michael Ziyang Diao, Krishna Balasubramanian, Sinho Chewi, and Adil Salim. Forward-backward Gaussian variational inference via JKO in the Bures–Wasserstein space. *International Conference on Machine Learning*, 2023.

Justin Domke. Provable smoothness guarantees for black-box variational inference. *International Conference on Machine Learning*, 2020.

Justin Domke, Robert Gower, and Guillaume Garrigos. Provable convergence guarantees for black-box variational inference. *Neural Information Processing Systems*, 2023.

Tomas Geffner and Justin Domke. Approximation based variance reduction for reparameterization gradients. *Neural Information Processing Systems*, 2020.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: Shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

Thomas Guilmeau, Hadrien Hendrikx, and Florence Forbes. Convergence of projected stochastic natural gradient variational inference for various step size and sample or batch size schedules. *International Conference on Artificial Intelligence and Statistics*, 2026.

D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

Matthew D. Hoffman and Yian Ma. Black-box variational inference as a parametric approximation to Langevin dynamics. *International Conference on Machine Learning*, 2020.

Yiheng Jiang, Sinho Chewi, and Aram-Alexandre Pooladian. Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space. *Conference on Learning Theory*, 2024.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Kyurae Kim, Jisu Oh, Kaiwen Wu, Yian Ma, and Jacob Gardner. On the convergence of black-box variational inference. *Neural Information Processing Systems*, 2023.

Kyurae Kim, Yian Ma, and Jacob Gardner. Linear convergence of black-box variational inference: Should we stick the landing? *International Conference on Artificial Intelligence and Statistics*, 2024.

Kyurae Kim, Qiang Fu, Yi-An Ma, Jacob R. Gardner, and Trevor Campbell. Stochastic gradient variational inference with Price's gradient estimator from Bures–Wasserstein to parameter space. *International Conference on Machine Learning*, 2026.

Navish Kumar, Thomas Möllenhoff, Mohammad Emtiyaz Khan, and Aurelien Lucchi. Optimization guarantees for square-root natural-gradient variational inference. *Transactions on Machine Learning Research*, 2025.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via Wasserstein gradient flows. *Neural Information Processing Systems*, 2022.

Jianfeng Lu and Yinchen Luo. Windowed thinning and query complexity for the bouncy particle and Zigzag samplers. *arXiv:2607.28413*, 2026.

Hoang Phuc Hau Luu, Hanlin Yu, Bernardo Williams, Marcelo Hartmann, and Arto Klami. Stochastic variance-reduced Gaussian variational inference on the Bures–Wasserstein manifold. *International Conference on Learning Representations*, 2025.

Yi-An Ma, Niladri S. Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L. Bartlett, and Michael I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3): 1942–1992, 2021.

Andrew C. Miller, Nicholas J. Foti, Alexander D'Amour, and Ryan P. Adams. Reducing reparameterization gradient variance. *Neural Information Processing Systems*, 2017.

Kevin Miller, Hao Li, and Andrea L. Bertozzi. Efficient graph-based active learning with probit likelihood via Gaussian approximations. *arXiv:2007.11126*, 2020.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

Jiyoung Park, Abhishek Roy, Jonathan W. Siegel, and Anirban Bhattacharya. Acceleration via silver step-size on Riemannian manifolds with applications to Wasserstein space. *Neural Information Processing Systems*, 2025.

Robert Price. A useful theorem for nonlinear devices having Gaussian inputs. *IRE Transactions on Information Theory*, 4(2):69–72, 1958.

Geoffrey Roeder, Yuhuai Wu, and David K. Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. *Neural Information Processing Systems*, 2017.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Fangyuan Sun, Ilyas Fatkhullin, and Niao He. Natural gradient VI: Guarantees for non-conjugate models. *Neural Information Processing Systems*, 2025.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. *International Conference on Artificial Intelligence and Statistics*, 2019.

Kaiwen Wu and Jacob R. Gardner. Understanding stochastic natural gradient variational inference. *International Conference on Machine Learning*, 2024.

A Omitted Proofs for Section 1 and Section 2

We first justify differentiating under the expectation. Recall

$$E(m, A) := \mathbb{E}_\xi[V(m + A\xi)], \qquad \xi \sim \mathcal{N}(0, I_d).$$

Assumption 1 gives $\|\nabla^2 V(x)\|_{\mathrm{op}} \le \beta$, and hence

$$\|\nabla V(x)\| \le \|\nabla V(0)\| + \beta\|x\|, \qquad |V(x)| \le |V(0)| + \|\nabla V(0)\|\|x\| + \frac{\beta}{2}\|x\|^2.$$

On any bounded neighborhood of (m, A), $V(m + A\xi)$ and its first two derivatives with respect to (m, A) are therefore bounded in absolute value by $C(1 + \|\xi\|^2)$ for some constant C. Since this bound is Gaussian-integrable, dominated convergence justifies differentiating twice under the expectation. We can now prove Lemma 1.

Proof of Lemma 1. Fix $\theta = (m, A)$ and a direction (u, U), and define

$$g(s) := E(m + su, A + sU) = \mathbb{E}_\xi \left[V\big(m + A\xi + s(u + U\xi)\big) \right].$$

The second derivative of the integrand is bounded in absolute value by $\beta\|u + U\xi\|^2$, which is integrable. Thus

$$g''(0) = \mathbb{E}_\xi \left[(u + U\xi)^\top \nabla^2 V(m + A\xi)(u + U\xi) \right].$$

Applying Assumption 1 pointwise gives

$$\alpha\mathbb{E}_\xi\|u + U\xi\|^2 \le g''(0) \le \beta\mathbb{E}_\xi\|u + U\xi\|^2.$$

Since $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\xi\xi^\top] = I_d$,

$$\mathbb{E}_\xi\|u + U\xi\|^2 = \|u\|^2 + \|U\|_F^2 = \|(u, U)\|^2.$$

Because $g''(0) = \nabla^2 E(\theta)[(u, U), (u, U)]$, the result follows. □

We then show the explicit proximal map of the entropy term used in Section 2.

Proposition 6. *For $\eta > 0$, let $\phi_\eta(y) := \frac{y + \sqrt{y^2 + 4\eta}}{2}$. If $\mathcal{A} = \mathbb{T}^d$, then $\mathrm{prox}_{\eta h}$ leaves the strictly lower-triangular entries unchanged and replaces each diagonal entry Y_{ii} by $\phi_\eta(Y_{ii})$. If $\mathcal{A} = \mathbb{S}^d$ and $Y = Q\,\mathrm{diag}(\lambda_1, \dots, \lambda_d)Q^\top$, then $\mathrm{prox}_{\eta h}(Y) = Q\,\mathrm{diag}\big(\phi_\eta(\lambda_1), \dots, \phi_\eta(\lambda_d)\big)Q^\top$.*

Proof. Suppose first that $\mathcal{A} = \mathbb{T}^d$. Since $-\log\det A = -\sum_i \log A_{ii}$, the proximal objective separates over the entries. Its strictly lower-triangular entries are minimized at $A_{ij} = Y_{ij}$, while each diagonal entry solves

$$\min_{\ell > 0} \left\{ -\log \ell + \frac{(\ell - Y_{ii})^2}{2\eta} \right\}.$$

The stationarity condition gives $\ell^2 - Y_{ii}\ell - \eta = 0$, whose unique positive solution is $\ell = \phi_\eta(Y_{ii})$.

Now suppose that $\mathcal{A} = \mathbb{S}^d$. The objective is strictly convex, coercive, and diverges at the boundary of $\mathbb{S}^d_{++}$, so it has a unique minimizer A^+. Its first-order condition is

$$-(A^+)^{-1} + \frac{A^+ - Y}{\eta} = 0,$$

and hence $Y = A^+ - \eta(A^+)^{-1}$. Thus, Y and A^+ are simultaneously diagonalizable. Writing

$$Y = Q\,\mathrm{diag}(\lambda_1, \dots, \lambda_d)Q^\top, \qquad A^+ = Q\,\mathrm{diag}(\ell_1, \dots, \ell_d)Q^\top,$$

we obtain $\ell_i^2 - \lambda_i\ell_i - \eta = 0$. Its unique positive solution is $\ell_i = \phi_\eta(\lambda_i)$, which proves the result. □

Now we state the precise accelerated proximal gradient method used in Section 2, then give the full convergence guarantee underlying Theorem 2.

Algorithm 3 Deterministic accelerated proximal gradient for GVI

Require: $x_0 = (m_0, A_0) \in \mathbb{R}^d \times \mathcal{A}_{++}$, $\eta = 1/\beta$, $\rho = (1 - \sqrt{\alpha/\beta})/(1 + \sqrt{\alpha/\beta})$, horizon T

$y_0 \leftarrow x_0$
for $k = 0, \dots, T-1$ **do**
  $g_k \leftarrow \nabla E(y_k)$
  $(m_{k+1}, \widetilde{A}_{k+1}) \leftarrow y_k - \eta g_k$
  $x_{k+1} \leftarrow (m_{k+1}, \mathrm{prox}_{\eta h}(\widetilde{A}_{k+1}))$
  $y_{k+1} \leftarrow x_{k+1} + \rho(x_{k+1} - x_k)$
end for
return $q_T = \mathcal{N}(m_T, A_T A_T^\top)$

Theorem 7 (Full version of Theorem 2)**.** *Let Assumption 1 hold, let $x_\star = (m_\star, A_\star)$ be the unique minimizer, and define*

$$\Delta_0 := F(q_0) - F^\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2.$$

For either factor representation $\mathcal{A} \in \{\mathbb{T}^d, \mathbb{S}^d\}$, every iterate generated by Algorithm 3 satisfies

$$F(q_k) - F^\star \le (1 - \kappa^{-1/2})^k \Delta_0$$

for every $k \ge 0$. Consequently, for every $0 < \varepsilon \le \Delta_0$, if the algorithm is run for

$$T_\varepsilon := \left\lceil \sqrt{\kappa} \log \frac{\Delta_0}{\varepsilon} \right\rceil$$

steps, then

$$F(q_{T_\varepsilon}) - F^\star \le \varepsilon.$$

The iteration complexity and exact energy-gradient oracle complexity are both

$$O\left(\sqrt{\kappa} \log \frac{\Delta_0}{\varepsilon}\right).$$

Proof. We first verify that every update is well defined. By Lemma 1, E is α-strongly convex and β-smooth on the entire linear space $\mathbb{R}^d \times \mathcal{A}$. Hence $\nabla E(y_k)$ is well defined even when the extrapolated factor in y_k does not belong to $\mathcal{A}_{++}$. Proposition 6 shows that the subsequent proximal step returns a factor in $\mathcal{A}_{++}$. Starting from $x_0 \in \mathbb{R}^d \times \mathcal{A}_{++}$, induction therefore shows that every accepted iterate x_k belongs to $\mathbb{R}^d \times \mathcal{A}_{++}$.

The function h is proper, closed, and convex. Therefore, the standard guarantee for accelerated proximal gradient with step size $1/\beta$ and momentum

$$\rho = \frac{1 - \sqrt{\alpha/\beta}}{1 + \sqrt{\alpha/\beta}}$$

gives (Nesterov, 2013)

$$F(x_k) - F(x_\star) \le (1 - \kappa^{-1/2})^k \left(F(x_0) - F(x_\star) + \frac{\alpha}{2}\|x_0 - x_\star\|^2\right).$$

For either $\mathcal{A} = \mathbb{T}^d$ or $\mathcal{A} = \mathbb{S}^d$, the map $A \mapsto AA^\top$ is a bijection from $\mathcal{A}_{++}$ onto $\mathbb{S}^d_{++}$. Thus minimizing in factor coordinates is equivalent to minimizing over $\mathcal{G}_d$, and

$$F(x_k) = F(q_k), \qquad F(x_\star) = F^\star.$$

Substitution into the preceding estimate proves the claimed contraction.

Finally, $1 - t \le e^{-t}$ gives

$$(1 - \kappa^{-1/2})^{T_\varepsilon} \Delta_0 \le e^{-T_\varepsilon/\sqrt{\kappa}} \Delta_0 \le \varepsilon.$$

Each update makes exactly one exact energy-gradient oracle call. Hence both the number of updates and the number of oracle calls equal T_ε, which proves the stated complexity bounds. □

Remark 2. Excluding the exact-expectation oracle call, a Cholesky iteration uses $O(d^2)$ factor arithmetic, whereas a symmetric-root iteration uses $O(d^3)$ arithmetic because of the eigendecomposition in the proximal map.

B PROOFS FOR SECTION 3

We now give the full construction and proof of Theorem 3. We first choose the cap size and dimension so that the standard Gaussian initialization has a constant objective gap while a query reveals the hidden cap with probability $\kappa^{-1+o(1)}$. We then smooth the construction so that all finite-order derivatives agree whenever the projection onto B stays away from the hidden cap, bound the probability of discovering the hidden direction, and finally show that without this information no fixed Gaussian can achieve a small objective gap for more than an exponentially small fraction of directions.

Throughout this section, all asymptotic notation is with respect to $\kappa \to \infty$. The symbols $c, C > 0$ denote constants whose values may change from line to line; the fixed constant C appearing in Theorem 3 will be chosen at the end of the proof. Each occurrence of "for all sufficiently large κ" means that the corresponding statement holds beyond a finite threshold depending only on constants fixed up to that point. At the end of the proof, κ_0 is chosen larger than all of the finitely many thresholds introduced.

Whenever probabilities are taken over the hidden direction v, it is uniformly distributed on $\mathbb{S}^{d-1}$ and independent of the algorithm's randomness.

For sufficiently large κ, define $R = R(\kappa) > 0$ as the unique solution of

$$\frac{\kappa - 1}{2}\mathbb{E}[(G - R)_+^2] = 1, \qquad G \sim \mathcal{N}(0, 1).$$

Indeed, $R \mapsto \mathbb{E}[(G - R)_+^2]$ is continuous and strictly decreasing,

$$\frac{\kappa - 1}{2}\mathbb{E}[(G - 0)_+^2] = \frac{\kappa - 1}{4} > 1$$

for all sufficiently large κ, while

$$\lim_{R \to \infty} \frac{\kappa - 1}{2}\mathbb{E}[(G - R)_+^2] = 0.$$

Hence the required positive solution exists and is unique. Moreover, this definition implies $R \to \infty$ as $\kappa \to \infty$. We use the standard Gaussian tail expansion

$$\mathbb{E}[(G - R)_+^2] = \frac{2}{\sqrt{2\pi}R^3} e^{-R^2/2} \left(1 + O(R^{-2})\right).$$

Substituting this into the defining equation for R gives

$$\frac{\kappa - 1}{\sqrt{2\pi}R^3} e^{-R^2/2} \left(1 + O(R^{-2})\right) = 1.$$

Taking logarithms yields

$$\frac{R^2}{2} = \log(\kappa - 1) - 3\log R - \frac{1}{2}\log(2\pi) + O(R^{-2}).$$

In particular, $R^2 = O(\log \kappa)$. Now, set

$$d := \lceil R^4 \rceil, \qquad a := (1 + R^{-1/2})\sqrt{d},$$

and let

$$B := \overline{B}_2(0, a) := \{x \in \mathbb{R}^d : \|x\| \le a\}$$

be the closed Euclidean ball of radius a centered at the origin. Then

$$d = \Theta(R^4) = \Theta((\log \kappa)^2), \qquad a - \sqrt{d} = R^{3/2} + o(R^{3/2}).$$

We write $\gamma := \mathcal{N}(0, I_d)$ for simplicity. For $v \in \mathbb{S}^{d-1}$, define

$$H_v := \{x : v^\top x \le R\}, \qquad K_v := B \cap H_v,$$

and, for every nonempty closed convex set K, let

$$e_K(x) := \frac{1}{2}\operatorname{dist}(x, K)^2.$$

Consider the potentials

$$V_v^0(x) := \frac{1}{2}\|x\|^2 + (\kappa - 1)e_{K_v}(x), \qquad V_B^0(x) := \frac{1}{2}\|x\|^2 + (\kappa - 1)e_B(x).$$

For every nonempty closed convex set K,

$$\nabla e_K(x) = x - \Pi_K x, \qquad \|\nabla e_K(x) - \nabla e_K(y)\| \le \|x - y\|.$$

Since e_K is convex and has a 1-Lipschitz gradient, both V_v^0 and V_B^0 are 1-strongly convex and have gradient Lipschitz constant at most κ. The following projection identities describe precisely where the hidden target V_v^0 agrees with the reference target V_B^0 and will be used repeatedly.

Lemma 8. *For every $x \in \mathbb{R}^d$,*

$$\Pi_{K_v} x = \Pi_B x \quad \Longleftrightarrow \quad v^\top \Pi_B x \le R.$$

Moreover,

$$\operatorname{dist}(x, H_v) \le \operatorname{dist}(x, K_v) \le \operatorname{dist}(x, H_v) + \operatorname{dist}(x, B).$$

Proof. Let $y = \Pi_B x$. If $v^\top y \le R$, then $y \in K_v$. Since $K_v \subseteq B$,

$$\operatorname{dist}(x, B) \le \operatorname{dist}(x, K_v) \le \|x - y\| = \operatorname{dist}(x, B).$$

Thus y attains the minimum distance from x to K_v, and uniqueness of the Euclidean projection onto the closed convex set K_v gives $\Pi_{K_v} x = y = \Pi_B x$. Conversely, if $\Pi_{K_v} x = \Pi_B x$, then $\Pi_B x \in K_v$, and hence $v^\top \Pi_B x \le R$.

The first distance inequality follows immediately from $K_v \subseteq H_v$. For the second, let $y = \Pi_{H_v} x$ and $z = \Pi_B y$. Since $0, y \in H_v$ and z lies on the line segment joining 0 and y, convexity of H_v gives $z \in H_v$. Hence $z \in B \cap H_v = K_v$. Therefore,

$$\operatorname{dist}(x, K_v) \le \|x - z\| \le \|x - y\| + \|y - z\| = \operatorname{dist}(x, H_v) + \operatorname{dist}(y, B).$$

Since $0 \in H_v$, nonexpansiveness of the projection onto H_v gives

$$\|y\| = \|\Pi_{H_v} x - \Pi_{H_v} 0\| \le \|x\|.$$

Because B is the ball of radius a centered at the origin,

$$\operatorname{dist}(y, B) = (\|y\| - a)_+ \le (\|x\| - a)_+ = \operatorname{dist}(x, B).$$

Combining the last two inequalities proves the lemma. □

We next show that the probability mass outside the enclosing ball is negligible.

Lemma 9. *If $X \sim \mathcal{N}(m, \Sigma)$, $\Sigma \preceq I_d$, and $\|m\| = o(R^{3/2})$, then*

$$\mathbb{E} \operatorname{dist}(X, B)^2 \le \exp\left[-\frac{1}{2}R^3 + o(R^3)\right].$$

Proof. Write $X = m + \Sigma^{1/2} G_d$ with $G_d \sim \mathcal{N}(0, I_d)$, and define $f(g) := \|m + \Sigma^{1/2} g\|$. Since $\Sigma \preceq I_d$,

$$|f(g) - f(h)| \le \|\Sigma^{1/2}(g - h)\| \le \|g - h\|,$$

so f is 1-Lipschitz. Moreover,

$$\mathbb{E}\|X\| \le \|m\| + \mathbb{E}\|\Sigma^{1/2} G_d\| \le \|m\| + \sqrt{\operatorname{tr} \Sigma} \le \|m\| + \sqrt{d}.$$

Hence, using $a = (1 + R^{-1/2})\sqrt{d}$, $d = \lceil R^4 \rceil$, and $\|m\| = o(R^{3/2})$,

$$a - \mathbb{E}\|X\| \ge R^{-1/2}\sqrt{d} - \|m\| = R^{3/2} + o(R^{3/2}).$$

Thus $a - \mathbb{E}\|X\| > 0$ for all sufficiently large κ. By Gaussian concentration for 1-Lipschitz functions,

$$\mathbb{P}\left(\|X\| - \mathbb{E}\|X\| \ge t\right) \le e^{-t^2/2}.$$

Since $\operatorname{dist}(X, B) = (\|X\| - a)_+$, tail integration gives

$$\mathbb{E}\operatorname{dist}(X, B)^2 = 2\int_0^\infty t\,\mathbb{P}(\|X\| > a + t)\,dt.$$

Therefore,

$$\mathbb{E}\operatorname{dist}(X, B)^2 \le 2\int_0^\infty t\exp\left[-\frac{1}{2}(a - \mathbb{E}\|X\| + t)^2\right] dt.$$

Using $(u + t)^2 \ge u^2 + t^2$ for $u, t \ge 0$,

$$\mathbb{E}\operatorname{dist}(X, B)^2 \le 2\exp\left[-\frac{1}{2}(a - \mathbb{E}\|X\|)^2\right].$$

The claimed estimate follows from

$$(a - \mathbb{E}\|X\|)^2 \ge R^3 + o(R^3).$$

□

We can now verify the $O(1)$ initial gap.

Proposition 10. *Uniformly over $v \in \mathbb{S}^{d-1}$,*

$$F_{V_v^0}(\gamma) - F^\star_{V_v^0} = 1 + o(1).$$

Proof. Since $\gamma = \mathcal{N}(0, I_d)$, for every $q \in \mathcal{G}_d$,

$$F_{V_v^0}(q) = \mathrm{KL}(q\,\|\,\gamma) + (\kappa - 1)\mathbb{E}_q[e_{K_v}(X)] - \frac{d}{2}\log(2\pi).$$

Consequently,

$$F_{V_v^0}(\gamma) - F^\star_{V_v^0} = (\kappa - 1)\mathbb{E}_\gamma[e_{K_v}(X)] - \inf_{p\in\mathcal{G}_d}\left\{\mathrm{KL}(p\,\|\,\gamma) + (\kappa - 1)\mathbb{E}_p[e_{K_v}(X)]\right\}.$$

We first evaluate the normalized objective at γ. Since $\|v\| = 1$ and $H_v = \{x : v^\top x \le R\}$,

$$\operatorname{dist}(x, H_v) = (v^\top x - R)_+.$$

Lemma 8 gives

$$0 \le \operatorname{dist}(x, K_v)^2 - \operatorname{dist}(x, H_v)^2 \le 2\operatorname{dist}(x, H_v)\operatorname{dist}(x, B) + \operatorname{dist}(x, B)^2.$$

For $X \sim \gamma$, rotational invariance gives $v^\top X \sim \mathcal{N}(0, 1)$, and hence

$$\mathbb{E}_\gamma\operatorname{dist}(X, H_v)^2 = \mathbb{E}[(G - R)_+^2] \le 1.$$

Therefore, by Cauchy–Schwarz,

$$0 \le \mathbb{E}_\gamma[e_{K_v}(X)] - \mathbb{E}_\gamma[e_{H_v}(X)] \le \left(\mathbb{E}_\gamma\operatorname{dist}(X, B)^2\right)^{1/2} + \frac{1}{2}\mathbb{E}_\gamma\operatorname{dist}(X, B)^2.$$

Lemma 9 gives

$$\mathbb{E}_\gamma[e_{K_v}(X)] - \mathbb{E}_\gamma[e_{H_v}(X)] \le \exp\left[-\frac{1}{4}R^3 + o(R^3)\right].$$

Since $\log\kappa = O(R^2) = o(R^3)$,

$$(\kappa - 1)\left(\mathbb{E}_\gamma[e_{K_v}(X)] - \mathbb{E}_\gamma[e_{H_v}(X)]\right) = O(e^{-cR^3}) = o(1)$$

uniformly over v. Using the defining equation for R,

$$(\kappa - 1)\mathbb{E}_\gamma[e_{K_v}(X)] = \frac{\kappa - 1}{2}\mathbb{E}[(G - R)_+^2] + o(1) = 1 + o(1)$$

uniformly over v.

It remains to show that the infimum of the normalized objective is $o(1)$. Set $\mu := \frac{\log R}{R}$ and consider the Gaussian $q := \mathcal{N}(-\mu v, I_d)$. Its relative entropy with respect to γ is

$$\mathrm{KL}(q \,\|\, \gamma) = \frac{\mu^2}{2} = O\left(\frac{(\log R)^2}{R^2}\right) = o(1).$$

Under q,

$$v^\top X \overset{d}{=} G - \mu, \qquad G \sim \mathcal{N}(0,1),$$

and hence

$$(\kappa - 1)\mathbb{E}_q[e_{H_v}(X)] = \frac{\kappa - 1}{2}\mathbb{E}[(G - (R + \mu))_+^2].$$

Using the defining relation for R,

$$(\kappa - 1)\mathbb{E}_q[e_{H_v}(X)] = \frac{\mathbb{E}[(G - (R + \mu))_+^2]}{\mathbb{E}[(G - R)_+^2]}.$$

The Gaussian tail expansion gives

$$\frac{\mathbb{E}[(G - (R + \mu))_+^2]}{\mathbb{E}[(G - R)_+^2]} = \left(\frac{R}{R + \mu}\right)^3 \exp\left[-R\mu - \frac{\mu^2}{2}\right](1 + o(1)).$$

Since $R\mu = \log R$ and $\mu/R \to 0$,

$$(\kappa - 1)\mathbb{E}_q[e_{H_v}(X)] = \frac{1 + o(1)}{R} = o(1).$$

Moreover, $\| - \mu v\| = o(R^{3/2})$, so Lemma 9 and the same comparison between K_v and H_v give

$$(\kappa - 1)\left(\mathbb{E}_q[e_{K_v}(X)] - \mathbb{E}_q[e_{H_v}(X)]\right) = o(1).$$

Therefore,

$$\mathrm{KL}(q \,\|\, \gamma) + (\kappa - 1)\mathbb{E}_q[e_{K_v}(X)] = o(1).$$

So,

$$0 \le \inf_{p \in \mathcal{G}_d} \{\mathrm{KL}(p \,\|\, \gamma) + (\kappa - 1)\mathbb{E}_p[e_{K_v}(X)]\} \le o(1).$$

□

We next smooth the construction to make the targets infinitely differentiable. Choose a nonnegative, even, radial density $\rho \in C_c^\infty(\mathbb{R}^d)$ supported on the closed unit ball $\overline{B}_2(0,1)$. Set

$$\tau := (\kappa - 1)^{-2}, \qquad \rho_\tau(u) := \tau^{-d}\rho(u/\tau), \qquad e_K^\tau := e_K * \rho_\tau.$$

Define

$$V_v(x) := \frac{1}{2}\|x\|^2 + (\kappa - 1)e_{K_v}^\tau(x), \qquad V_B(x) := \frac{1}{2}\|x\|^2 + (\kappa - 1)e_B^\tau(x).$$

Lemma 11. *For every nonempty closed convex set K,*

$$0 \le e_K^\tau(x) - e_K(x) \le \frac{\tau^2}{2}.$$

Moreover,

$$0 \preceq \nabla^2 e_K^\tau(x) \preceq I_d.$$

Consequently, $V_v \in C^\infty(\mathbb{R}^d)$ and

$$I_d \preceq \nabla^2 V_v(x) \preceq \kappa I_d.$$

If

$$v^\top \Pi_B x \le R - \tau,$$

then V_v and V_B have the same derivatives of every finite order at x.

Furthermore, for every $q \in \mathcal{G}_d$,

$$0 \le F_{V_v}(q) - F_{V_v^0}(q) \le \frac{1}{2(\kappa - 1)^3},$$

and

$$0 \le F^\star_{V_v} - F^\star_{V_v^0} \le \frac{1}{2(\kappa-1)^3}.$$

In particular,

$$F_{V_v}(\gamma) - F^\star_{V_v} = 1 + o(1)$$

uniformly over $v \in \mathbb{S}^{d-1}$.

Proof. Let $Z \sim \rho_\tau$. Since ρ_τ is even, $\mathbb{E}Z = 0$. Hence, by convexity of e_K and Jensen's inequality,

$$e_K^\tau(x) = \mathbb{E}[e_K(x - Z)] \ge e_K(x - \mathbb{E}Z) = e_K(x).$$

Since e_K has a 1-Lipschitz gradient,

$$e_K(x - Z) \le e_K(x) - \langle \nabla e_K(x), Z\rangle + \frac{1}{2}\|Z\|^2.$$

Taking expectations and using $\mathbb{E}Z = 0$ and $\|Z\| \le \tau$ gives

$$0 \le e_K^\tau(x) - e_K(x) \le \frac{\tau^2}{2}.$$

Since $\rho_\tau \in C_c^\infty(\mathbb{R}^d)$, convolution with ρ_τ gives $e_K^\tau \in C^\infty(\mathbb{R}^d)$. Moreover, e_K^τ is convex because it is an average of translates of the convex function e_K. Also,

$$\nabla e_K^\tau(x) = \mathbb{E}[\nabla e_K(x - Z)],$$

and therefore

$$\|\nabla e_K^\tau(x) - \nabla e_K^\tau(y)\| \le \mathbb{E}\|\nabla e_K(x - Z) - \nabla e_K(y - Z)\| \le \|x - y\|.$$

Thus e_K^τ is smooth and convex with a 1-Lipschitz gradient, so $0 \preceq \nabla^2 e_K^\tau(x) \preceq I_d$. The Hessian bounds for V_v follow immediately.

The map $x \mapsto v^\top \Pi_B x$ is 1-Lipschitz because Π_B is nonexpansive and $\|v\| = 1$. Hence, if

$$v^\top \Pi_B x \le R - \tau$$

and $\|u\| \le \tau$, then

$$v^\top \Pi_B(x - u) \le v^\top \Pi_B x + \|u\| \le R.$$

By Lemma 8, this implies

$$\Pi_{K_v}(x - u) = \Pi_B(x - u),$$

and therefore

$$e_{K_v}(x - u) = e_B(x - u).$$

For every multi-index α, differentiating the convolution kernel gives

$$\partial^\alpha(e_{K_v}^\tau - e_B^\tau)(x) = \int_{\mathbb{R}^d} [e_{K_v}(x - u) - e_B(x - u)]\, \partial^\alpha \rho_\tau(u)\, du = 0,$$

since $\operatorname{supp}(\partial^\alpha \rho_\tau) \subseteq \overline{B}_2(0, \tau)$. Thus V_v and V_B have the same values and all finite-order derivatives at x, including when $v^\top \Pi_B x = R - \tau$.

Finally, the first bound and $\tau = (\kappa - 1)^{-2}$ give, for every $x \in \mathbb{R}^d$,

$$0 \le V_v(x) - V_v^0(x) \le \frac{\kappa - 1}{2}\tau^2 = \frac{1}{2(\kappa-1)^3}.$$

Taking expectations under any $q \in \mathcal{G}_d$ gives

$$0 \le F_{V_v}(q) - F_{V_v^0}(q) \le \frac{1}{2(\kappa-1)^3}.$$

Taking infima over $q \in \mathcal{G}_d$ similarly yields

$$0 \le F^\star_{V_v} - F^\star_{V_v^0} \le \frac{1}{2(\kappa-1)^3}.$$

Therefore,

$$\left|\left(F_{V_v}(\gamma) - F^\star_{V_v}\right) - \left(F_{V_v^0}(\gamma) - F^\star_{V_v^0}\right)\right| \le \frac{1}{2(\kappa-1)^3} = o(1).$$

Proposition 10 now gives

$$F_{V_v}(\gamma) - F^\star_{V_v} = 1 + o(1)$$

uniformly over v. □

We next bound the probability that a point query reaches the region where the oracle response depends on v.

Lemma 12. *There exists a constant $C > 0$ such that, for every fixed $z \in \mathbb{R}^d$ and $0 < u < \|z\|$,*

$$\mathbb{P}_v(v^\top z > u) \le \frac{C\|z\|}{u\sqrt{d}} \exp\left[-\frac{d-1}{2}\frac{u^2}{\|z\|^2}\right].$$

Moreover, for all sufficiently large κ, uniformly over $x \in \mathbb{R}^d$,

$$\mathbb{P}_v\left(v^\top \Pi_B x > R - \tau\right) \le \frac{C}{\kappa} \exp\left[C(\log\kappa)^{3/4}\right].$$

Proof. By rotational invariance, we may assume without loss of generality that $z = \|z\|e_1$. Writing $v = (v_1, \dots, v_d)$, we then have $v^\top z = \|z\|v_1$. To derive its distribution, let

$$g = (g_1, \dots, g_d) \sim \mathcal{N}(0, I_d).$$

Since the standard Gaussian distribution is rotationally invariant, $v := \frac{g}{\|g\|}$ is uniformly distributed on $\mathbb{S}^{d-1}$. In particular,

$$v_1^2 = \frac{g_1^2}{g_1^2 + \cdots + g_d^2}.$$

Set

$$U := g_1^2, \qquad W := g_2^2 + \cdots + g_d^2.$$

Then U and W are independent and satisfy

$$U \sim \chi_1^2, \qquad W \sim \chi_{d-1}^2.$$

Equivalently, they are independent Gamma random variables with the same scale parameter and shape parameters $1/2$ and $(d-1)/2$, respectively. Therefore,

$$\frac{U}{U+W} \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{d-1}{2}\right).$$

Thus $T := v_1^2$ has density

$$f_T(t) = \frac{1}{B\left(\frac{1}{2}, \frac{d-1}{2}\right)} t^{-1/2}(1-t)^{(d-3)/2}, \qquad 0 < t < 1,$$

where $B(a,b) := \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function. The distribution of v_1 is symmetric about zero. Since $T = v_1^2$, a change of variables gives, for $0 < x < 1$,

$$f_T(x^2) = \frac{f_{v_1}(x) + f_{v_1}(-x)}{2x} = \frac{f_{v_1}(x)}{x}.$$

Hence

$$f_{v_1}(x) = c_d(1-x^2)^{(d-3)/2}, \qquad -1 < x < 1,$$

where $c_d := \frac{1}{B\left(\frac{1}{2}, \frac{d-1}{2}\right)}$. Using $\Gamma(1/2) = \sqrt{\pi}$, we obtain

$$c_d = \frac{\Gamma(d/2)}{\sqrt{\pi}\,\Gamma((d-1)/2)}.$$

Finally, using Stirling's formula gives

$$\frac{\Gamma(x+1/2)}{\Gamma(x)} = x^{1/2}\left(1 + O(x^{-1})\right).$$

Taking $x = (d-1)/2$ yields

$$c_d = \sqrt{\frac{d-1}{2\pi}}\left(1 + O(d^{-1})\right).$$

so there exists a constant $C > 0$ such that $c_d \le C\sqrt{d}$ for every $d - 1 \ge 1$.

Since

$$\mathbb{P}_v(v^\top z > u) = c_d \int_{u/\|z\|}^1 (1-t^2)^{(d-3)/2}\, dt,$$

and $t \geq u/\|z\|$ on the domain of integration,

$$\int_{u/\|z\|}^1 (1-t^2)^{(d-3)/2}\, dt \leq \frac{\|z\|}{u} \int_{u/\|z\|}^1 t(1-t^2)^{(d-3)/2}\, dt = \frac{\|z\|}{u(d-1)} \left(1 - \frac{u^2}{\|z\|^2}\right)^{(d-1)/2}.$$

Using $c_d \leq C\sqrt{d}$ and, for $d \geq 2$,

$$\frac{\sqrt{d}}{d-1} = \frac{1}{\sqrt{d}} \frac{d}{d-1} \leq \frac{2}{\sqrt{d}},$$

we obtain

$$\mathbb{P}_v(v^\top z > u) \leq \frac{C\|z\|}{u\sqrt{d}} \left(1 - \frac{u^2}{\|z\|^2}\right)^{(d-1)/2}.$$

Finally, since $1 - x \leq e^{-x}$ for $x \geq 0$,

$$\mathbb{P}_v(v^\top z > u) \leq \frac{C\|z\|}{u\sqrt{d}} \exp\left[-\frac{d-1}{2}\frac{u^2}{\|z\|^2}\right].$$

We apply the preceding estimate with $z = \Pi_B x$ and $u = R - \tau$. If

$$\|\Pi_B x\| \leq R - \tau,$$

then $v^\top \Pi_B x \leq \|\Pi_B x\| \leq R - \tau$ for every $v \in \mathbb{S}^{d-1}$, so the probability is zero. Otherwise, since $\Pi_B x \in B$ and B is the Euclidean ball of radius a,

$$\|\Pi_B x\| \leq a = (1 + R^{-1/2})\sqrt{d}.$$

the preceding bound gives

$$\mathbb{P}_v\left(v^\top \Pi_B x > R - \tau\right) \leq C \exp\left[-\frac{d-1}{d}\frac{(R-\tau)^2}{2(1+R^{-1/2})^2}\right],$$

after increasing the constant C if necessary. Since $d = \lceil R^4 \rceil$,

$$\frac{d-1}{d} = 1 + O(R^{-4}).$$

Moreover, $\tau = (\kappa - 1)^{-2}$ is negligible compared with any inverse power of R, so

$$(R-\tau)^2 = R^2 + o(1).$$

Using

$$(1 + R^{-1/2})^{-2} = 1 - 2R^{-1/2} + 3R^{-1} + O(R^{-3/2}),$$

we obtain

$$\frac{d-1}{d}\frac{(R-\tau)^2}{2(1+R^{-1/2})^2} = \frac{R^2}{2} - R^{3/2} + O(R).$$

Also,

$$\frac{R^2}{2} = \log\kappa - \frac{3}{2}\log\log\kappa + O(1),$$

so $R^2 = O(\log\kappa)$ and hence

$$R^{3/2} = O((\log\kappa)^{3/4}).$$

Increasing C once more if necessary therefore gives

$$\mathbb{P}_v\left(v^\top \Pi_B x > R - \tau\right) \leq \frac{C}{\kappa} \exp\left[C(\log\kappa)^{3/4}\right]$$

uniformly over $x \in \mathbb{R}^d$. □

We next compare the algorithm with its trajectory under V_B. Until the hidden region is reached, the two trajectories coincide.

Lemma 13. *Fix an integer $r \geq 1$ and a randomized algorithm making at most N pointwise queries. At each query point, the algorithm observes the value of V and its derivatives up to order r. Let $x_1^0, \dots, x_N^0$ denote its queries against V_B (ignoring unused indices if the algorithm terminates early). If*

$$v^\top \Pi_B x_t^0 \leq R - \tau$$

for every query made by the algorithm, then the oracle responses, query points, and final outputs are identical for V_v and V_B. Moreover,

$$\mathbb{P}_v\left(\exists t : v^\top \Pi_B x_t^0 > R - \tau\right) \leq C\frac{N}{\kappa}\exp\left[C(\log\kappa)^{3/4}\right].$$

Proof. Condition on the algorithm's randomness $x_t^0 = x_t^0(\omega)$, the first query is then the same under V_v and V_B. Suppose that the two transcripts agree through query $t-1$. Since the algorithm chooses its next query from the previous transcript, the t-th query is also the same and equals x_t^0. If

$$v^\top \Pi_B x_t^0 \leq R - \tau,$$

Lemma 11 implies that V_v and V_B return the same derivatives through order r at x_t^0. Hence the transcripts continue to agree. By induction, the two query sequences, oracle responses, and final outputs are identical until a query satisfies

$$v^\top \Pi_B x_t^0 > R - \tau.$$

The points $x_1^0, \dots, x_N^0$ are determined by V_B and therefore do not depend on v. Applying Lemma 12 to each point and taking a union bound gives

$$\mathbb{P}_v\left(\exists t : v^\top \Pi_B x_t^0 > R - \tau\right) \leq C\frac{N}{\kappa}\exp\left[C(\log\kappa)^{3/4}\right].$$

Averaging over the algorithm's randomness preserves the same bound. □

We now show that if the hidden cap is not discovered, the resulting output Gaussian can achieve a small objective gap for only an exponentially small fraction of directions.

Lemma 14. *There exist constants $\varepsilon_0, c > 0$ such that, for all sufficiently large κ and every fixed Gaussian $q \in \mathcal{G}_d$,*

$$\mathbb{P}_v\left(F_{V_v}(q) - F_{V_v}^\star \leq 2\varepsilon_0\right) \leq e^{-cR^2}.$$

Proof. Choose $\varepsilon_0 > 0$ sufficiently small to satisfy the conditions below, and then take κ sufficiently large. All choices will be independent of q.

The proof of Proposition 10 and Lemma 11 imply that

$$\inf_{p \in \mathcal{G}_d}\left\{\mathrm{KL}(p\,\|\,\gamma) + (\kappa-1)\mathbb{E}_p[e_{K_v}^\tau(X)]\right\} = o(1)$$

uniformly over v. Hence, if $F_{V_v}(q) - F_{V_v}^\star \leq 2\varepsilon_0$, then, for all sufficiently large κ,

$$\mathrm{KL}(q\,\|\,\gamma) + (\kappa-1)\mathbb{E}_q[e_{K_v}^\tau(X)] \leq 3\varepsilon_0.$$

Since $e_{K_v}^\tau \geq e_{K_v} \geq e_{H_v}$, we obtain

$$\mathrm{KL}(q\,\|\,\gamma) + \frac{\kappa-1}{2}\mathbb{E}_q[(v^\top X - R)_+^2] \leq 3\varepsilon_0.$$

Write $q = \mathcal{N}(m, \Sigma)$, and let $\lambda_1, \dots, \lambda_d$ be the eigenvalues of Σ. The Gaussian KL formula yields

$$\|m\|^2 + \sum_{i=1}^d(\lambda_i - \log\lambda_i - 1) \leq 6\varepsilon_0.$$

Each summand is nonnegative and is bounded away from zero outside $[1/2, 2]$. Moreover,

$$\lambda - \log\lambda - 1 \geq \frac{1}{8}(\lambda-1)^2, \qquad \lambda \in [1/2, 2].$$

Thus our choice of sufficiently small ε_0 ensures

$$\|m\|^2 \le 6\varepsilon_0, \qquad \frac{1}{2} I_d \preceq \Sigma \preceq 2 I_d,$$

and

$$\|I_d - \Sigma\|_F^2 = \sum_{i=1}^{d} (\lambda_i - 1)^2 \le 48\varepsilon_0.$$

Set

$$\sigma^2 := v^\top \Sigma v, \qquad t := \frac{R - v^\top m}{\sigma}.$$

Since $v^\top X \overset{d}{=} v^\top m + \sigma G$, $G \sim \mathcal{N}(0,1)$, we have

$$\mathbb{E}_q[(v^\top X - R)_+^2] = \sigma^2 \mathbb{E}[(G - t)_+^2].$$

Hence the preceding bound gives

$$\frac{\kappa - 1}{2} \sigma^2 \mathbb{E}[(G - t)_+^2] \le 3\varepsilon_0.$$

Using $\frac{\kappa-1}{2}\mathbb{E}[(G - R)_+^2] = 1$, we obtain

$$\sigma^2 \frac{\mathbb{E}[(G - t)_+^2]}{\mathbb{E}[(G - R)_+^2]} \le 3\varepsilon_0.$$

Since $\sigma^2 \in [1/2, 2]$ and $|v^\top m| \le \sqrt{6\varepsilon_0}$, we have $t/R \in [1/2, 2]$ for all sufficiently large κ. The two-sided Gaussian tail estimate

$$c \frac{e^{-u^2/2}}{u^3} \le \mathbb{E}[(G - u)_+^2] \le C \frac{e^{-u^2/2}}{u^3}, \qquad u \ge 1,$$

therefore gives

$$c \exp\left[-\frac{t^2 - R^2}{2}\right] \le \sigma^2 \frac{\mathbb{E}[(G - t)_+^2]}{\mathbb{E}[(G - R)_+^2]} \le 3\varepsilon_0.$$

Choosing ε_0 sufficiently small ensures $t^2 - R^2 \ge 2$. We have

$$\sigma^2(t^2 - R^2) = (R - v^\top m)^2 - \sigma^2 R^2.$$

Expanding the square and using $\|v\| = 1$ gives

$$\sigma^2(t^2 - R^2) = R^2 v^\top (I_d - \Sigma) v - 2R v^\top m + (v^\top m)^2.$$

Hence

$$-2R v^\top m + R^2 v^\top (I_d - \Sigma) v = \sigma^2(t^2 - R^2) - (v^\top m)^2.$$

Using $t^2 - R^2 \ge 2$, $\sigma^2 \ge 1/2$, and $(v^\top m)^2 \le \|m\|^2 \le 6\varepsilon_0$, we obtain

$$-2R v^\top m + R^2 v^\top (I_d - \Sigma) v \ge 1 - 6\varepsilon_0 \ge \frac{1}{2}$$

by choosing $\varepsilon_0 \le 1/12$. Thus, if the objective gap is at most $2\varepsilon_0$, then at least one of $v^\top(-m) \ge \frac{1}{8R}$ or $v^\top (I_d - \Sigma) v \ge \frac{1}{4R^2}$ must hold. We show that each event has probability at most e^{-cR^2} over the random direction v.

For the first event, if $\|m\| < 1/(8R)$, then it is impossible. Otherwise, Lemma 12 with $z = -m$ and $u = 1/(8R)$ gives

$$\mathbb{P}_v\left(v^\top(-m) \ge \frac{1}{8R}\right) \le \frac{CR\|m\|}{\sqrt{d}} \exp\left[-c \frac{d}{R^2 \|m\|^2}\right].$$

Since $\|m\|^2 \le 6\varepsilon_0$ and $d = \lceil R^4 \rceil$,

$$\frac{d}{R^2 \|m\|^2} \ge cR^2,$$

while $\frac{CR\|m\|}{\sqrt{d}}$ is also bounded for all sufficiently large κ. Hence

$$\mathbb{P}_v\left(v^\top(-m) \geq \frac{1}{8R}\right) \leq Ce^{-cR^2}.$$

For the second event, write $v = \frac{g}{\|g\|}$, $g \sim \mathcal{N}(0, I_d)$. On the event $\{\|g\|^2 \geq d/2\}$,

$$v^\top(I_d - \Sigma)v \geq \frac{1}{4R^2}$$

implies

$$g^\top(I_d - \Sigma)g \geq \frac{d}{8R^2}.$$

Moreover,

$$|\operatorname{tr}(I_d - \Sigma)| \leq \sqrt{d}\,\|I_d - \Sigma\|_F \leq \sqrt{48\varepsilon_0}\sqrt{d}.$$

Since $d = \lceil R^4 \rceil$, by choosing ε_0 sufficiently small, we get $|\operatorname{tr}(I_d - \Sigma)| \leq \frac{d}{16R^2}$. Therefore,

$$g^\top(I_d - \Sigma)g - \operatorname{tr}(I_d - \Sigma) \geq \frac{d}{16R^2}.$$

Now the Hanson–Wright inequality (Hanson & Wright, 1971) gives

$$\mathbb{P}\left(g^\top(I_d - \Sigma)g - \operatorname{tr}(I_d - \Sigma) \geq u\right) \leq 2\exp\left[-c\min\left\{\frac{u^2}{\|I_d - \Sigma\|_F^2}, \frac{u}{\|I_d - \Sigma\|_{\mathrm{op}}}\right\}\right].$$

Using

$$\|I_d - \Sigma\|_F \leq \sqrt{48\varepsilon_0}, \qquad \|I_d - \Sigma\|_{\mathrm{op}} \leq \sqrt{48\varepsilon_0},$$

and taking $u = d/(16R^2)$, together with $d = \lceil R^4 \rceil$, gives

$$\mathbb{P}\left(g^\top(I_d - \Sigma)g - \operatorname{tr}(I_d - \Sigma) \geq \frac{d}{16R^2}\right) \leq 2e^{-cR^2}$$

for a constant $c > 0$. Moreover, $\|g\|^2 \sim \chi^2_d$, and the standard chi-square lower-tail bound gives

$$\mathbb{P}(\|g\|^2 < d/2) \leq e^{-cd} \leq e^{-cR^2}$$

for all sufficiently large κ. The last inequality holds since $d \geq R^4$. So, the event

$$F_{V_v}(q) - F^\star_{V_v} \leq 2\varepsilon_0$$

is therefore contained in the union of these two events. Hence

$$\mathbb{P}_v\left(F_{V_v}(q) - F^\star_{V_v} \leq 2\varepsilon_0\right) \leq Ce^{-cR^2}.$$

Since C is fixed, decreasing $c > 0$ and increasing the threshold on κ if necessary gives

$$\mathbb{P}_v\left(F_{V_v}(q) - F^\star_{V_v} \leq 2\varepsilon_0\right) \leq e^{-cR^2}.$$

□

We now combine Lemmas 13 and 14 to prove the main theorem.

Proof of Theorem 3. Fix the constants $\varepsilon_0, c > 0$ supplied by Lemma 14, and let $C_1 \geq 1$ be a fixed constant for which the probability estimate in Lemma 13 holds. We first establish the required bounds for all sufficiently large κ and then choose a single threshold κ_0.

Fix an integer $r \geq 1$ and a randomized order-r algorithm that makes at most N queries. For a fixed algorithm randomness ω, let $x^0_1, \ldots, x^0_N$ denote its queries against V_B, and let $q^0(\omega)$ be the corresponding output. By Lemma 13,

$$\mathbb{P}_v\left(\exists t : v^\top \Pi_B x^0_t > R - \tau\right) \leq C_1\frac{N}{\kappa}\exp\left[C_1(\log\kappa)^{3/4}\right].$$

Run the algorithm against V_v using the same ω. If none of the reference queries satisfies $v^\top \Pi_B x_t^0 > R - \tau$, the two transcripts and outputs coincide, so the actual output is $q = q^0(\omega)$.

For each fixed ω, the reference output $q^0(\omega)$ does not depend on v. Lemma 14 therefore gives

$$\mathbb{P}_v\left(F_{V_v}(q^0(\omega)) - F_{V_v}^\star \le 2\varepsilon_0\right) \le e^{-cR^2}.$$

Thus, if the actual output has objective gap at most $2\varepsilon_0$, either a reference query reaches the hidden cap or the reference output itself has objective gap at most $2\varepsilon_0$. Taking a union bound and averaging over the algorithm's randomness yields

$$\mathbb{P}_{v,\mathrm{alg}}\left(F_{V_v}(q) - F_{V_v}^\star \le 2\varepsilon_0\right) \le C_1 \frac{N}{\kappa} \exp\left[C_1(\log\kappa)^{3/4}\right] + e^{-cR^2}.$$

Now suppose that the algorithm satisfies

$$\sup_{v\in\mathbb{S}^{d-1}} \mathbb{E}\left[F_{V_v}(q) - F_{V_v}^\star\right] \le \varepsilon_0.$$

Averaging over v gives

$$\mathbb{E}_{v,\mathrm{alg}}\left[F_{V_v}(q) - F_{V_v}^\star\right] \le \varepsilon_0.$$

Since the objective gap is nonnegative, Markov's inequality implies

$$\mathbb{P}_{v,\mathrm{alg}}\left(F_{V_v}(q) - F_{V_v}^\star \le 2\varepsilon_0\right) \ge \frac{1}{2}.$$

Combining the two probability bounds gives

$$\frac{1}{2} \le C_1 \frac{N}{\kappa} \exp\left[C_1(\log\kappa)^{3/4}\right] + e^{-cR^2}.$$

Because $R \to \infty$, the last term is at most $1/4$ for all sufficiently large κ. Hence

$$C_1 \frac{N}{\kappa} \exp\left[C_1(\log\kappa)^{3/4}\right] \ge \frac{1}{4},$$

and therefore

$$N \ge \frac{\kappa}{4C_1} \exp\left[-C_1(\log\kappa)^{3/4}\right].$$

To remove the fixed multiplicative factor, choose the constant in the theorem as $C := C_1 + 1$. Since for all sufficiently large κ,

$$\exp\left[-(\log\kappa)^{3/4}\right] \le \frac{1}{4C_1}.$$

Consequently,

$$N \ge \frac{\kappa}{4C_1} \exp\left[-C_1(\log\kappa)^{3/4}\right] \ge \kappa \exp\left[-C(\log\kappa)^{3/4}\right].$$

It remains to verify the dimension and initial-gap requirements. The construction gives

$$d = \Theta((\log\kappa)^2),$$

and Lemma 11 gives

$$V_v \in C^\infty(\mathbb{R}^d), \qquad I_d \preceq \nabla^2 V_v(x) \preceq \kappa I_d$$

for every $v \in \mathbb{S}^{d-1}$ and $x \in \mathbb{R}^d$. Thus the entire family belongs to $\mathcal{V}_d(\kappa)$. The same lemma also gives

$$F_{V_v}(\gamma) - F_{V_v}^\star = 1 + o(1)$$

uniformly over $v \in \mathbb{S}^{d-1}$. Fix $C_0 := 2$. Then, for all sufficiently large κ,

$$\sup_{v\in\mathbb{S}^{d-1}} \left(F_{V_v}(\gamma) - F_{V_v}^\star\right) \le C_0.$$

We have now fixed ε_0, C_0, and C. Choose $\kappa_0 > 1$ larger than the finitely many thresholds needed for the construction and the preceding lemmas, and this proves the theorem. □

C PROOFS FOR SECTION 4

We first analyze the Cholesky finite-difference construction and its variants, and then turn to the symmetric-root score estimators.

C.1 CHOLESKY PROXIMAL ACCELERATION

We first establish the unbiasedness and variance bound of the finite-difference estimator used in Algorithm 1.

Proposition 15. *For every* $(m, L) \in \mathbb{R}^d \times \mathbb{T}^d$,

$$\mathbb{E}[G^{\mathrm{FD}}(m, L; \xi)] = \nabla E(m, L).$$

Under Assumption 2,

$$\mathbb{E}\|G^{\mathrm{FD}}(m, L; \xi) - \nabla E(m, L)\|^2 \le \frac{L_H^2}{2}\left(3(d+5)\|L\|_F^4 + \nu^2 d(d+3)\|L\|_F^2\right).$$

If V is quadratic, then $G^{\mathrm{FD}}(m, L; \xi) = \nabla E(m, L)$ for every ξ.

Proof. Fix (m, L) and write $D = D_\nu(m)$ and $r = r_\nu(m, L; \xi)$. Since $\mathbb{E}\xi = 0$ and $\mathbb{E}[\xi\xi^\top] = I_d$,

$$\mathbb{E}[\nabla V(m) + r] = \mathbb{E}[\nabla V(m + L\xi)], \qquad DL + \mathbb{E}[r\xi^\top] = \mathbb{E}[\nabla V(m + L\xi)\xi^\top].$$

Taking the lower-triangular part proves unbiasedness. Let $H = \nabla^2 V(m)$. For each j,

$$D_\nu(m; e_j) - He_j = \int_0^1 [\nabla^2 V(m + t\nu e_j) - H]e_j \, dt.$$

Therefore

$$\|D - H\|_{\mathrm{op}} \le \|D - H\|_F \le \frac{L_H|\nu|\sqrt{d}}{2}.$$

Taylor's theorem gives

$$\|\nabla V(m + L\xi) - \nabla V(m) - HL\xi\| \le \frac{L_H}{2}\|L\xi\|^2,$$

and hence

$$\|r\| \le \frac{L_H}{2}\|L\xi\|^2 + \frac{L_H|\nu|\sqrt{d}}{2}\|L\xi\|.$$

Since $\mathbb{E}G^{\mathrm{FD}} = \nabla E$, the centered estimator is

$$G^{\mathrm{FD}} - \nabla E = \left(r - \mathbb{E}r, \mathrm{low}\left(r\xi^\top - \mathbb{E}[r\xi^\top]\right)\right).$$

For every square-integrable Z,

$$\mathbb{E}\|Z - \mathbb{E}Z\|^2 \le \mathbb{E}\|Z\|^2.$$

Moreover, $\mathrm{low}(\cdot)$ is a Frobenius contraction and $\|r\xi^\top\|_F^2 = \|r\|^2\|\xi\|^2$. Hence

$$\mathbb{E}\|G^{\mathrm{FD}} - \nabla E\|^2 \le \mathbb{E}[(1 + \|\xi\|^2)\|r\|^2].$$

Using $(a+b)^2 \le 2a^2 + 2b^2$,

$$\|r\|^2 \le \frac{L_H^2}{2}\left(\|L\xi\|^4 + \nu^2 d\|L\xi\|^2\right).$$

Let $A = L^\top L \succeq 0$. Then $\|L\xi\|^2 = \xi^\top A\xi$ and $\mathrm{tr}(A) = \|L\|_F^2$. The Gaussian identities

$$\mathbb{E}[\xi_i\xi_j] = \delta_{ij}, \qquad \mathbb{E}[\xi_i\xi_j\xi_k\xi_\ell] = \delta_{ij}\delta_{k\ell} + \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}$$

give

$$\mathbb{E}[\xi^\top A\xi] = \mathrm{tr}(A), \qquad \mathbb{E}[(\xi^\top A\xi)^2] = (\mathrm{tr}\, A)^2 + 2\,\mathrm{tr}(A^2) \le 3(\mathrm{tr}\, A)^2$$

and

$$\mathbb{E}[\|\xi\|^2(\xi^\top A\xi)] = (d+2)\,\mathrm{tr}(A).$$

Consequently,

$$\mathbb{E}[(1+\|\xi\|^2)\|L\xi\|^2] = (d+3)\|L\|_F^2.$$

Writing $\xi = Ru$ with $R = \|\xi\|$ and using $\mathbb{E}[R^6] = (d+4)\mathbb{E}[R^4]$ gives

$$\mathbb{E}[\|\xi\|^2\|L\xi\|^4] = (d+4)\mathbb{E}\|L\xi\|^4 \le 3(d+4)\|L\|_F^4.$$

Thus

$$\mathbb{E}[(1+\|\xi\|^2)\|L\xi\|^4] \le 3(d+5)\|L\|_F^4.$$

Combining the estimates yields

$$\mathbb{E}\|G^{\mathrm{FD}} - \nabla E\|^2 \le \frac{L_H^2}{2}\left(3(d+5)\|L\|_F^4 + \nu^2 d(d+3)\|L\|_F^2\right).$$

If V is quadratic with Hessian H, then $D_\nu(m; e_j) = He_j$ for every j and $\nabla V(m+L\xi) = \nabla V(m) + HL\xi$. Thus $D = H$, $r = 0$, and $G^{\mathrm{FD}}(m, L; \xi) = \nabla E(m, L)$ for every ξ. □

We next derive the constrained proximal map used in the Cholesky update.

Proposition 16. *For $\eta > 0$ and $Y \in \mathbb{T}^d$, define $L^{(\gamma)}$ for $\gamma \ge 0$ by*

$$L_{ij}^{(\gamma)} := \frac{Y_{ij}}{1+\gamma} \quad (i > j), \qquad L_{ii}^{(\gamma)} := \frac{Y_{ii} + \sqrt{Y_{ii}^2 + 4\eta(1+\gamma)}}{2(1+\gamma)}.$$

If $\|L^{(0)}\|_F \le \sqrt{d/\alpha}$, then $\mathrm{prox}_{\eta\bar{h}}(Y) = L^{(0)}$. Otherwise, there is a unique $\gamma > 0$ satisfying $\|L^{(\gamma)}\|_F^2 = d/\alpha$ and $\mathrm{prox}_{\eta\bar{h}}(Y) = L^{(\gamma)}$.

Proof. The proximal subproblem is

$$\min_{L \in \mathbb{T}_{++}^d} \left\{ -\eta \log\det L + \frac{1}{2}\|L - Y\|_F^2 : \|L\|_F^2 \le \frac{d}{\alpha} \right\}.$$

The objective is strictly convex and diverges as any diagonal entry approaches zero, so the minimizer is unique and belongs to $\mathbb{T}_{++}^d$. Let $\gamma \ge 0$ be the multiplier for the norm constraint. The KKT conditions are

$$(1+\gamma)L_{ij} = Y_{ij} \quad (i > j), \qquad (1+\gamma)L_{ii} - Y_{ii} - \frac{\eta}{L_{ii}} = 0.$$

The diagonal equation has a unique positive root, giving $L = L^{(\gamma)}$. If $\|L^{(0)}\|_F^2 \le d/\alpha$, complementary slackness gives $\gamma = 0$. Otherwise, the norm constraint is active and $\|L^{(\gamma)}\|_F^2 = d/\alpha$. For $i > j$, $|L_{ij}^{(\gamma)}| = |Y_{ij}|/(1+\gamma)$ is nonincreasing in γ. For each diagonal entry, differentiating the stationarity equation gives

$$\frac{dL_{ii}^{(\gamma)}}{d\gamma} = -\frac{L_{ii}^{(\gamma)}}{1 + \gamma + \eta/(L_{ii}^{(\gamma)})^2} < 0.$$

Hence $\gamma \mapsto \|L^{(\gamma)}\|_F$ is continuous and strictly decreasing. Moreover,

$$\lim_{\gamma \to \infty} \|L^{(\gamma)}\|_F = 0.$$

Since $\|L^{(0)}\|_F > \sqrt{d/\alpha}$, there is a unique $\gamma > 0$ satisfying the constraint. □

We now prove the fixed-batch proximal acceleration result used in the complexity analysis.

Lemma 17. *Let $\Phi = f + h$, where f is α-strongly convex and β-smooth and h is proper, closed, and convex. Let $x_\star$ minimize Φ, write $\Phi^\star := \Phi(x_\star)$ and $\kappa := \beta/\alpha$, and set*

$$s := \sqrt{\max\{\kappa, 4\}}, \qquad \eta := \frac{1}{\alpha s^2}, \qquad \rho := \frac{1 - s^{-1}}{1 + s^{-1}}.$$

Suppose $x_0 = y_0$ and

$$x_{k+1} = \mathrm{prox}_{\eta h}(y_k - \eta g_k), \qquad y_{k+1} = x_{k+1} + \rho(x_{k+1} - x_k),$$

where, conditionally on the past $\mathcal{F}_k$,

$$\mathbb{E}[g_k \mid \mathcal{F}_k] = \nabla f(y_k), \qquad \mathbb{E}[\|g_k - \nabla f(y_k)\|^2 \mid \mathcal{F}_k] \le \sigma^2.$$

Then, for every integer $T \ge 0$,

$$\mathbb{E}[\Phi(x_T) - \Phi^\star] \le e^{-T/s}\left(\Phi(x_0) - \Phi^\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2\right) + \frac{\sigma^2}{\alpha s}.$$

Proof. Since $s \ge 2$ and $s^2 \ge \kappa = \beta/\alpha$, we have $\beta\eta \le 1$. Proximal optimality gives

$$p_k := \frac{y_k - x_{k+1}}{\eta} - g_k \in \partial h(x_{k+1}), \qquad \widetilde{g}_k := g_k + p_k = \frac{y_k - x_{k+1}}{\eta}.$$

Define

$$\Lambda_k(z) := f(y_k) + \langle g_k, z - y_k\rangle + \frac{\alpha}{2}\|z - y_k\|^2 + h(x_{k+1}) + \langle p_k, z - x_{k+1}\rangle.$$

By strong convexity of f, convexity of h, and conditional unbiasedness,

$$\mathbb{E}[\Lambda_k(z) \mid \mathcal{F}_k] \le \Phi(z)$$

for every $\mathcal{F}_k$-measurable z. Let

$$\bar{x}_k := \mathrm{prox}_{\eta h}(y_k - \eta\nabla f(y_k)).$$

Decomposing $x_{k+1} - y_k = (x_{k+1} - \bar{x}_k) + (\bar{x}_k - y_k)$ and using conditional centering gives

$$\mathbb{E}[\langle \nabla f(y_k) - g_k, x_{k+1} - y_k\rangle \mid \mathcal{F}_k] = \mathbb{E}[\langle \nabla f(y_k) - g_k, x_{k+1} - \bar{x}_k\rangle \mid \mathcal{F}_k].$$

Nonexpansiveness of the proximal map gives

$$\|x_{k+1} - \bar{x}_k\| \le \eta\|g_k - \nabla f(y_k)\|,$$

and therefore

$$\mathbb{E}[\langle \nabla f(y_k) - g_k, x_{k+1} - y_k\rangle \mid \mathcal{F}_k] \le \eta\sigma^2.$$

By β-smoothness,

$$\Phi(x_{k+1}) \le f(y_k) + h(x_{k+1}) + \langle \nabla f(y_k), x_{k+1} - y_k\rangle + \frac{\beta}{2}\|x_{k+1} - y_k\|^2.$$

Since

$$\Lambda_k(y_k) = f(y_k) + h(x_{k+1}) + \langle p_k, y_k - x_{k+1}\rangle,$$

we obtain

$$\Phi(x_{k+1}) \le \Lambda_k(y_k) + \langle \widetilde{g}_k, x_{k+1} - y_k\rangle + \langle \nabla f(y_k) - g_k, x_{k+1} - y_k\rangle + \frac{\beta}{2}\|x_{k+1} - y_k\|^2.$$

Using $x_{k+1} - y_k = -\eta\widetilde{g}_k$ and taking conditional expectations,

$$\mathbb{E}[\Phi(x_{k+1}) \mid \mathcal{F}_k] \le \mathbb{E}[\Lambda_k(y_k) \mid \mathcal{F}_k] + \left(\frac{\beta\eta^2}{2} - \eta\right)\mathbb{E}[\|\widetilde{g}_k\|^2 \mid \mathcal{F}_k] + \eta\sigma^2.$$

Define

$$\varphi_0(z) := \Phi(x_0) + \frac{\alpha}{2}\|z - x_0\|^2, \qquad \varphi_{k+1} := (1 - s^{-1})\varphi_k + s^{-1}\Lambda_k.$$

Let

$$v_k := \arg\min_z \varphi_k(z), \qquad \varphi_k^{\min} := \varphi_k(v_k).$$

Every φ_k is a quadratic with Hessian αI, so

$$\varphi_k(z) = \varphi_k^{\min} + \frac{\alpha}{2}\|z - v_k\|^2.$$

The stationarity condition for φ_{k+1} gives

$$v_{k+1} = (1 - s^{-1})v_k + s^{-1}y_k - \frac{1}{\alpha s}\widetilde{g}_k.$$

We next establish

$$y_k = \frac{x_k + s^{-1}v_k}{1 + s^{-1}}.$$

It holds at $k = 0$. Assuming it at iteration k and using $x_{k+1} = y_k - \eta\widetilde{g}_k$ gives

$$s^{-1}v_{k+1} = x_{k+1} - (1 - s^{-1})x_k,$$

which is equivalent to the momentum update for y_{k+1}. Set

$$\zeta_0 := 0, \qquad \zeta_{k+1} := (1 - s^{-1})\zeta_k + \eta\sigma^2.$$

We prove by induction that

$$\mathbb{E}[\varphi_k^{\min}] \geq \mathbb{E}[\Phi(x_k)] - \zeta_k.$$

The claim holds with equality at $k = 0$. Since

$$x_k - y_k = -s^{-1}(v_k - y_k),$$

the lower-model property at x_k gives

$$\Lambda_k(x_k) = \Lambda_k(y_k) - s^{-1}\langle\widetilde{g}_k, v_k - y_k\rangle + \frac{\alpha}{2s^2}\|v_k - y_k\|^2.$$

Completing the square in φ_{k+1} gives

$$\varphi_{k+1}^{\min} = (1-s^{-1})\varphi_k^{\min} + s^{-1}\Lambda_k(y_k) + (1-s^{-1})s^{-1}\left(\langle\widetilde{g}_k, v_k - y_k\rangle + \frac{\alpha}{2}\|v_k - y_k\|^2\right) - \frac{1}{2\alpha s^2}\|\widetilde{g}_k\|^2.$$

Using the induction hypothesis and the lower-model bound at x_k, the inner-product terms cancel and yield

$$\mathbb{E}[\varphi_{k+1}^{\min}] \geq \mathbb{E}[\Lambda_k(y_k)] - (1 - s^{-1})\zeta_k - \frac{\eta}{2}\mathbb{E}\|\widetilde{g}_k\|^2 + \frac{\alpha}{2}(1 - s^{-1})s^{-1}(1 + s^{-1})\mathbb{E}\|v_k - y_k\|^2.$$

Dropping the final nonnegative term,

$$\mathbb{E}[\varphi_{k+1}^{\min}] \geq \mathbb{E}[\Lambda_k(y_k)] - (1 - s^{-1})\zeta_k - \frac{\eta}{2}\mathbb{E}\|\widetilde{g}_k\|^2.$$

The one-step upper bound implies

$$\mathbb{E}[\Lambda_k(y_k)] \geq \mathbb{E}[\Phi(x_{k+1})] + \left(\eta - \frac{\beta\eta^2}{2}\right)\mathbb{E}\|\widetilde{g}_k\|^2 - \eta\sigma^2.$$

Therefore

$$\mathbb{E}[\varphi_{k+1}^{\min}] \geq \mathbb{E}[\Phi(x_{k+1})] - \zeta_{k+1} + \frac{\eta}{2}(1 - \beta\eta)\mathbb{E}\|\widetilde{g}_k\|^2.$$

Since $\beta\eta \leq 1$, the induction closes. Finally, the lower-model property at $x_\star$ gives

$$\mathbb{E}[\varphi_T(x_\star)] - \Phi^\star \leq (1 - s^{-1})^T\left(\varphi_0(x_\star) - \Phi^\star\right).$$

Using the bound established above,

$$\mathbb{E}[\varphi_T^{\min}] \geq \mathbb{E}[\Phi(x_T)] - \zeta_T,$$

together with $\varphi_T^{\min} \leq \varphi_T(x_\star)$, we obtain

$$\mathbb{E}[\Phi(x_T) - \Phi^\star] \leq (1 - s^{-1})^T\left(\Phi(x_0) - \Phi^\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2\right) + \zeta_T.$$

Moreover,

$$\zeta_T = \eta\sigma^2\sum_{j=0}^{T-1}(1 - s^{-1})^j \leq \eta s\sigma^2 = \frac{\sigma^2}{\alpha s}.$$

Using $(1 - s^{-1})^T \leq e^{-T/s}$ completes the proof. □

We can now combine the estimator bound with the acceleration lemma.

Theorem 18 (Full version of Theorem 4)**.** *Suppose Assumptions 1 and 2 hold,* $x_0 \in \operatorname{dom} \bar{h}$*,* $0 < |\nu| \le \alpha^{-1/2}$*, and* $0 < \varepsilon \le \Delta_0$*, where*

$$\Delta_0 := F(q_0) - F^\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2.$$

Let

$$s := \sqrt{\max\{\kappa, 4\}}, \qquad b := \left\lceil \max\left\{1, \frac{1494\lambda_H^2 d^3}{s\varepsilon}\right\}\right\rceil, \qquad T_\varepsilon := \left\lceil s \log \frac{2\Delta_0}{\varepsilon}\right\rceil.$$

Choose

$$\eta := \frac{1}{\alpha s^2}, \qquad \rho := \frac{s-1}{s+1}.$$

Then Algorithm 1 with G_b^{FD} *satisfies*

$$\mathbb{E}[F(q_{T_\varepsilon}) - F^\star] \le \varepsilon.$$

The exact first-order oracle count is

$$N_{\nabla V} = T_\varepsilon(d + 1 + b),$$

and consequently

$$T_\varepsilon = O\left(\sqrt{\kappa}\log\frac{2\Delta_0}{\varepsilon}\right), \qquad N_{\nabla V} = O\left(\left(d\sqrt{\kappa} + \frac{\lambda_H^2 d^3}{\varepsilon}\right)\log\frac{2\Delta_0}{\varepsilon}\right).$$

Proof. Write $x_k = (m_k, L_k)$ and $y_k = (m_k^y, L_k^y)$. Every proximal iterate satisfies $\|L_k\|_F \le \sqrt{d/\alpha}$. Since $L_k^y = (1+\rho)L_k - \rho L_{k-1}$ for $k \ge 1$ and $0 \le \rho < 1$, $\|L_k^y\|_F < 3\sqrt{d/\alpha}$, and hence

$$\|L_k^y\|_F^2 < \frac{9d}{\alpha}, \qquad \|L_k^y\|_F^4 < \frac{81d^2}{\alpha^2}.$$

Proposition 15 and $|\nu| \le \alpha^{-1/2}$ give

$$\mathbb{E}[G^{\mathrm{FD}}(m_k^y, L_k^y; \xi) \mid \mathcal{F}_k] = \nabla E(m_k^y, L_k^y)$$

and

$$\mathbb{E}[\|G^{\mathrm{FD}}(m_k^y, L_k^y; \xi) - \nabla E(m_k^y, L_k^y)\|^2 \mid \mathcal{F}_k] \le \frac{L_H^2}{\alpha^2}(126d + 621)d^2.$$

Let

$$\sigma^2 := \frac{L_H^2}{\alpha^2}(126d + 621)d^2.$$

For G_b^{FD},

$$\mathbb{E}[G_b^{\mathrm{FD}}(y_k) \mid \mathcal{F}_k] = \nabla E(y_k), \qquad \mathbb{E}[\|G_b^{\mathrm{FD}}(y_k) - \nabla E(y_k)\|^2 \mid \mathcal{F}_k] \le \frac{\sigma^2}{b}.$$

Using $\lambda_H = L_H/\alpha^{3/2}$,

$$\frac{\sigma^2}{\alpha} = (126d + 621)\lambda_H^2 d^2 \le 747\lambda_H^2 d^3.$$

Define $\Phi(m, L) := E(m, L) + \bar{h}(L)$. The norm restriction in $\bar{h}$ does not change the optimizer, and every accepted iterate lies in its domain. Therefore

$$\Phi(x_\star) = F^\star, \qquad \Phi(x_k) = F(q_k).$$

Applying Lemma 17 with variance parameter σ^2/b gives

$$\mathbb{E}[F(q_{T_\varepsilon}) - F^\star] \le e^{-T_\varepsilon/s}\Delta_0 + \frac{\sigma^2}{\alpha b s} \le e^{-T_\varepsilon/s}\Delta_0 + \frac{747\lambda_H^2 d^3}{bs}.$$

By construction,

$$e^{-T_\varepsilon/s}\Delta_0 \le \frac{\varepsilon}{2}, \qquad \frac{747\lambda_H^2 d^3}{bs} \le \frac{\varepsilon}{2}.$$

Thus the desired accuracy holds. Each update constructs $D_\nu(m_k^y)$ using $d+1$ gradient evaluations and uses b additional evaluations for the batch. Hence

$$N_{\nabla V} = T_\varepsilon(d + 1 + b),$$

which yields the stated complexity. □

The same accelerated update also admits a second-order estimator with improved dimension dependence.

Proposition 19. *Under Assumption 2, for*

$$G^{\mathrm{paired}}(m, L; \xi, \zeta) := \left(\nabla V(m + L\xi) - \nabla^2 V(m + L\zeta)L\xi, \mathrm{low}(\nabla^2 V(m + L\zeta)L)\right),$$

where $\xi, \zeta \sim \mathcal{N}(0, I_d)$ are independent,

$$\mathbb{E}[G^{\mathrm{paired}}(m, L; \xi, \zeta)] = \nabla E(m, L),$$

and

$$\mathbb{E}\|G^{\mathrm{paired}}(m, L; \xi, \zeta) - \nabla E(m, L)\|^2 \le \frac{9}{2} L_H^2 \|L\|_F^4.$$

If V is quadratic, then $G^{\mathrm{paired}}(m, L; \xi, \zeta) = \nabla E(m, L)$ for every ξ, ζ.

Proof. Let

$$X := m + L\xi, \qquad Y := m + L\zeta, \qquad H_Y := \nabla^2 V(Y).$$

Since ξ and ζ are independent and $\mathbb{E}\xi = 0$, we have $\mathbb{E}[H_Y L\xi] = 0$. So, the mean component has expectation $\nabla_m E(m, L)$. For the covariance-factor component, Stein's identity gives

$$\mathbb{E}[\nabla V(m + L\zeta)\zeta^\top] = \mathbb{E}[\nabla^2 V(m + L\zeta)L],$$

and therefore

$$\mathrm{low}\left(\mathbb{E}[H_Y L]\right) = \nabla_L E(m, L).$$

This proves unbiasedness. Let

$$H := \nabla^2 V(m), \qquad z := L\xi, \qquad y := L\zeta.$$

Since the mean minimizes mean-square distance,

$$\mathbb{E}\|G^{\mathrm{paired}} - \nabla E\|^2 \le \mathbb{E}\left\|G^{\mathrm{paired}} - (\nabla V(m), \mathrm{low}(HL))\right\|^2.$$

For the first component,

$$\nabla V(m+z) - \nabla V(m) - \nabla^2 V(m+y)z = (\nabla V(m + z) - \nabla V(m) - Hz) + \left(H - \nabla^2 V(m + y)\right) z.$$

Hessian Lipschitzness gives

$$\|\nabla V(m + z) - \nabla V(m) - Hz\| \le \frac{L_H}{2}\|z\|^2$$

and

$$\|(H - \nabla^2 V(m + y))z\| \le L_H \|y\|\|z\|.$$

Thus

$$\|\nabla V(m + z) - \nabla V(m) - \nabla^2 V(m + y)z\|^2 \le \frac{L_H^2}{2}\|z\|^4 + 2L_H^2\|y\|^2\|z\|^2.$$

Independence gives $\mathbb{E}[\|y\|^2\|z\|^2] = \|L\|_F^4$, while $\mathbb{E}\|L\xi\|^4 \le 3\|L\|_F^4$. Hence

$$\mathbb{E}\|\nabla V(X) - H_Y L\xi - \nabla V(m)\|^2 \le \frac{7}{2} L_H^2 \|L\|_F^4.$$

For the covariance-factor component,

$$\|\,\mathrm{low}((H_Y - H)L)\|_F \le \|(H_Y - H)L\|_F \le L_H\|y\|\|L\|_F,$$

and therefore

$$\mathbb{E}\|\,\mathrm{low}(H_Y L) - \mathrm{low}(HL)\|_F^2 \le L_H^2 \|L\|_F^4.$$

Adding the two components gives the result. If V is quadratic, its Hessian is constant and both errors vanish. □

Combining this variance bound with Lemma 17 gives the second-order complexity.

Corollary 20. *Suppose Assumptions 1 and 2 hold, $x_0 \in \operatorname{dom} \bar{h}$, and $0 < \varepsilon \le \Delta_0$, where*

$$\Delta_0 := F(q_0) - F^\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2.$$

Let

$$s := \sqrt{\max\{\kappa, 4\}}, \qquad b := \left\lceil \max\left\{1, \frac{729\lambda_H^2 d^2}{s\varepsilon}\right\}\right\rceil, \qquad T_\varepsilon := \left\lceil s \log \frac{2\Delta_0}{\varepsilon}\right\rceil.$$

Choose

$$\eta := \frac{1}{\alpha s^2}, \qquad \rho := \frac{s-1}{s+1}.$$

Run the update of Algorithm 1 with G_b^{paired} in place of G_b^{FD}. Then

$$\mathbb{E}[F(q_{T_\varepsilon}) - F^\star] \le \varepsilon.$$

The exact oracle counts are

$$N_{\nabla V} = bT_\varepsilon, \qquad N_{\mathrm{HVP}} = dbT_\varepsilon.$$

Consequently,

$$N_{\nabla V} = O\left(\left(\sqrt{\kappa} + \frac{\lambda_H^2 d^2}{\varepsilon}\right)\log\frac{2\Delta_0}{\varepsilon}\right),$$

and

$$N_{\mathrm{HVP}} = O\left(\left(d\sqrt{\kappa} + \frac{\lambda_H^2 d^3}{\varepsilon}\right)\log\frac{2\Delta_0}{\varepsilon}\right).$$

Proof. As in the proof of Theorem 18, $\|L_k^y\|_F < 3\sqrt{d/\alpha}$, so $\|L_k^y\|_F^4 < 81d^2/\alpha^2$. Proposition 19 gives

$$\mathbb{E}[G_b^{\mathrm{paired}}(y_k) \mid \mathcal{F}_k] = \nabla E(y_k)$$

and

$$\mathbb{E}[\|G_b^{\mathrm{paired}}(y_k) - \nabla E(y_k)\|^2 \mid \mathcal{F}_k] \le \frac{\sigma_2^2}{b}, \qquad \sigma_2^2 := \frac{729}{2}\frac{L_H^2 d^2}{\alpha^2}.$$

Thus

$$\frac{\sigma_2^2}{\alpha} = \frac{729}{2}\lambda_H^2 d^2.$$

Applying Lemma 17,

$$\mathbb{E}[F(q_{T_\varepsilon}) - F^\star] \le e^{-T_\varepsilon/s}\Delta_0 + \frac{729\lambda_H^2 d^2}{2bs}.$$

The prescribed choices make both terms at most $\varepsilon/2$. Each realization of G^{paired} uses one gradient evaluation and d HVPs, giving the stated oracle counts. □

C.2 Cholesky projected acceleration

The finite-difference estimator can alternatively be combined with an exact entropy derivative and Euclidean projection. Define

$$K_L := \left\{(m, L) : m \in \mathbb{R}^d,\ L \in \mathbb{T}^d,\ L_{ii} \ge \frac{1}{2\sqrt{\beta}},\ \|L\|_F \le \sqrt{\frac{d}{\alpha}}\right\}.$$

To extend the entropy gradient outside K_L, define

$$\psi(t) := \begin{cases} \log(2\sqrt{\beta}) - 2\sqrt{\beta}\left(t - \dfrac{1}{2\sqrt{\beta}}\right) + 2\beta\left(t - \dfrac{1}{2\sqrt{\beta}}\right)^2, & t < \dfrac{1}{2\sqrt{\beta}}, \\ -\log t, & t \ge \dfrac{1}{2\sqrt{\beta}}. \end{cases} \tag{2}$$

and

$$\widetilde{h}(L) := \sum_{i=1}^d \psi(L_{ii}), \qquad \widetilde{F}(m, L) := E(m, L) + \widetilde{h}(L).$$

The next lemma establishes the properties needed for projected acceleration.

Lemma 21. *The minimizer $x_\star = (m_\star, L_\star)$ belongs to K_L. Moreover, $\widetilde{F}$ is globally α-strongly convex and 5β-smooth on $\mathbb{R}^d \times \mathbb{T}^d$, and*

$$\widetilde{F} = F, \qquad \nabla\widetilde{F} = \nabla F \qquad \text{on } K_L.$$

Proof. Since $\Sigma_\star = L_\star L_\star^\top \preceq \alpha^{-1} I$, $\|L_\star\|_F^2 = \operatorname{tr}(\Sigma_\star) \le d/\alpha$. Let $\Sigma_{\star,[i]}$ denote the leading $i \times i$ principal submatrix of $\Sigma_\star$. The Cholesky factorization gives

$$L_{\star,ii}^2 = \frac{1}{(\Sigma_{\star,[i]}^{-1})_{ii}}.$$

Since $\Sigma_{\star,[i]} \succeq \beta^{-1} I$, we have $\Sigma_{\star,[i]}^{-1} \preceq \beta I$, and hence $L_{\star,ii} \ge \frac{1}{\sqrt{\beta}} > \frac{1}{2\sqrt{\beta}}$. Therefore $x_\star \in K_L$.

The function ψ is continuously differentiable and convex, with

$$\psi''(t) = \begin{cases} 4\beta, & t < \dfrac{1}{2\sqrt{\beta}}, \\ \dfrac{1}{t^2}, & t > \dfrac{1}{2\sqrt{\beta}}. \end{cases}$$

Thus $0 \le \psi''(t) \le 4\beta$, so $\widetilde{h}$ is convex and 4β-smooth. By Lemma 1, E is α-strongly convex and β-smooth. Hence $\widetilde{F}$ is globally α-strongly convex and 5β-smooth.

For $(m, L) \in K_L$, all diagonal entries satisfy $L_{ii} \ge 1/(2\sqrt{\beta})$, so

$$\widetilde{h}(L) = -\sum_{i=1}^{d} \log L_{ii} = -\log\det L.$$

The two branches of ψ have the same first derivative at $1/(2\sqrt{\beta})$, so the gradients also agree. □

The projection itself also admits a simple one-dimensional characterization.

Proposition 22. *For $(m, Y) \in \mathbb{R}^d \times \mathbb{T}^d$, the Euclidean projection onto K_L is*

$$P_{K_L}(m, Y) = (m, L^{(\gamma)}),$$

where

$$L_{ij}^{(\gamma)} := \frac{Y_{ij}}{1+\gamma} \quad (i > j), \qquad L_{ii}^{(\gamma)} := \max\left\{\frac{1}{2\sqrt{\beta}}, \frac{Y_{ii}}{1+\gamma}\right\}.$$

If $\|L^{(0)}\|_F \le \sqrt{d/\alpha}$, then $\gamma = 0$. Otherwise, $\gamma > 0$ is the unique solution of $\|L^{(\gamma)}\|_F^2 = d/\alpha$.

Proof. The mean coordinate is unconstrained. For the factor coordinate, consider

$$\min_{L \in \mathbb{T}^d} \left\{\frac{1}{2}\|L - Y\|_F^2 : L_{ii} \ge \frac{1}{2\sqrt{\beta}},\ \|L\|_F^2 \le \frac{d}{\alpha}\right\}.$$

Let $\gamma \ge 0$ be the multiplier for the norm constraint and $\mu_i \ge 0$ the multipliers for the diagonal lower bounds. The KKT conditions give

$$(1+\gamma)L_{ij} = Y_{ij} \quad (i > j), \qquad (1+\gamma)L_{ii} = Y_{ii} + \mu_i,$$

with

$$\mu_i \left(L_{ii} - \frac{1}{2\sqrt{\beta}}\right) = 0.$$

Thus

$$L_{ii} = \max\left\{\frac{1}{2\sqrt{\beta}}, \frac{Y_{ii}}{1+\gamma}\right\}.$$

If $\|L^{(0)}\|_F \le \sqrt{d/\alpha}$, the norm constraint is inactive. Otherwise it is active. The map $\gamma \mapsto \|L^{(\gamma)}\|_F$ is continuous and nonincreasing and is strictly decreasing while its value exceeds $\sqrt{d/\alpha}$. Moreover,

$$\lim_{\gamma \to \infty} \|L^{(\gamma)}\|_F = \frac{\sqrt{d}}{2\sqrt{\beta}} < \sqrt{\frac{d}{\alpha}}.$$

Hence there is a unique positive solution of the active constraint. □

Algorithm 4 Cholesky projected acceleration

Require: $x_0 \in K_L$, batch size b, estimator $G_b \in \{G_b^{\mathrm{FD}}, G_b^{\mathrm{paired}}\}$, FD scale ν, horizon T

1: $y_0 \leftarrow x_0$
2: **for** $k = 0, \dots, T-1$ **do**
3: $\quad \widehat{g}_k \leftarrow G_b(y_k)$
4: $\quad$ write $y_k = (m_k^y, L_k^y)$ and set $g_k \leftarrow \widehat{g}_k + (0, \nabla\widetilde{h}(L_k^y))$
5: $\quad x_{k+1} \leftarrow P_{K_L}(y_k - \eta g_k)$
6: $\quad y_{k+1} \leftarrow x_{k+1} + \rho(x_{k+1} - x_k)$
7: **end for**
8: **return** $q_T = \mathcal{N}(m_T, L_T L_T^\top)$

These ingredients lead to Algorithm 4, which is a projected counterpart of the proximal method.

The extrapolated point y_k need not be feasible, but both E and $\widetilde{h}$ are defined on the entire ambient space. The projection ensures that every accepted iterate belongs to K_L.

The same estimator bounds therefore give the following projected guarantee.

Theorem 23. *Suppose Assumptions 1 and 2 hold, $x_0 \in K_L$, $0 < |\nu| \le \alpha^{-1/2}$, and $0 < \varepsilon \le \Delta_0$, where*

$$\Delta_0 := F(q_0) - F^\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2.$$

Let

$$s := \sqrt{\max\{5\kappa, 4\}}, \qquad b := \left\lceil \max\left\{1, \frac{1494\lambda_H^2 d^3}{s\varepsilon}\right\}\right\rceil, \qquad T_\varepsilon := \left\lceil s \log\frac{2\Delta_0}{\varepsilon}\right\rceil.$$

Then Algorithm 4 with G_b^{FD} satisfies

$$\mathbb{E}[F(q_{T_\varepsilon}) - F^\star] \le \varepsilon.$$

The exact first-order oracle count is

$$N_{\nabla V} = T_\varepsilon(d + 1 + b),$$

and consequently

$$N_{\nabla V} = O\left(\left(d\sqrt{\kappa} + \frac{\lambda_H^2 d^3}{\varepsilon}\right)\log\frac{2\Delta_0}{\varepsilon}\right).$$

Proof. Every accepted iterate belongs to K_L, and the momentum relation gives $\|L_k^y\|_F < 3\sqrt{\frac{d}{\alpha}}$. Proposition 15 therefore gives

$$\mathbb{E}[G_b^{\mathrm{FD}}(y_k) \mid \mathcal{F}_k] = \nabla E(y_k), \qquad \mathbb{E}[\|G_b^{\mathrm{FD}}(y_k) - \nabla E(y_k)\|^2 \mid \mathcal{F}_k] \le \frac{\sigma^2}{b},$$

where $\sigma^2/\alpha \le 747\lambda_H^2 d^3$. Let

$$g_k := G_b^{\mathrm{FD}}(y_k) + (0, \nabla\widetilde{h}(L_k^y)).$$

Since the entropy gradient is deterministic,

$$\mathbb{E}[g_k \mid \mathcal{F}_k] = \nabla\widetilde{F}(y_k)$$

with the same stochastic error. By Lemma 21, $\widetilde{F}$ is α-strongly convex and 5β-smooth. Moreover,

$$P_{K_L} = \mathrm{prox}_{\eta\iota_{K_L}}.$$

Thus Lemma 17 applies with $f = \widetilde{F}$, $h = \iota_{K_L}$, condition number 5κ, and variance parameter σ^2/b. Hence

$$\mathbb{E}[F(q_{T_\varepsilon}) - F^\star] \le e^{-T_\varepsilon/s}\Delta_0 + \frac{747\lambda_H^2 d^3}{bs}.$$

The prescribed T_ε and b make the two terms at most $\varepsilon/2$. The oracle count is the same as in the proximal finite-difference method. □

The second order estimator yields the analogous result.

Corollary 24. *Suppose Assumptions 1 and 2 hold, $x_0 \in K_L$, and $0 < \varepsilon \le \Delta_0$, where*

$$\Delta_0 := F(q_0) - F^\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2.$$

Let

$$s := \sqrt{\max\{5\kappa, 4\}}, \qquad b := \left\lceil \max\left\{1, \frac{729\lambda_H^2 d^2}{s\varepsilon}\right\}\right\rceil, \qquad T_\varepsilon := \left\lceil s \log\frac{2\Delta_0}{\varepsilon}\right\rceil.$$

Run Algorithm 4 with G_b^{paired} in place of G_b^{FD}. Then

$$\mathbb{E}[F(q_{T_\varepsilon}) - F^\star] \le \varepsilon.$$

The exact oracle counts are

$$N_{\nabla V} = bT_\varepsilon, \qquad N_{\mathrm{HVP}} = dbT_\varepsilon.$$

Consequently,

$$N_{\nabla V} = O\left(\left(\sqrt{\kappa} + \frac{\lambda_H^2 d^2}{\varepsilon}\right)\log\frac{2\Delta_0}{\varepsilon}\right),$$

and

$$N_{\mathrm{HVP}} = O\left(\left(d\sqrt{\kappa} + \frac{\lambda_H^2 d^3}{\varepsilon}\right)\log\frac{2\Delta_0}{\varepsilon}\right).$$

Proof. As above, $\|L_k^y\|_F < 3\sqrt{d/\alpha}$. Proposition 19 gives

$$\mathbb{E}[G_b^{\mathrm{paired}}(y_k) \mid \mathcal{F}_k] = \nabla E(y_k)$$

and

$$\mathbb{E}[\|G_b^{\mathrm{paired}}(y_k) - \nabla E(y_k)\|^2 \mid \mathcal{F}_k] \le \frac{\sigma^2}{b}, \qquad \sigma^2 := \frac{729}{2}\frac{L_H^2 d^2}{\alpha^2}.$$

Adding the deterministic entropy gradient preserves this stochastic error. Lemma 17 gives

$$\mathbb{E}[F(q_{T_\varepsilon}) - F^\star] \le e^{-T_\varepsilon/s}\Delta_0 + \frac{729\lambda_H^2 d^2}{2bs}.$$

The parameter choices make both terms at most $\varepsilon/2$. Each realization uses one gradient evaluation and d HVPs. □

C.3 Symmetric-root score acceleration

We begin by recording the gradient identities underlying STL and Price–STL. Let

$$X := m + C\xi, \qquad \xi \sim \mathcal{N}(0, I_d),$$

with $C \in \mathbb{S}_{++}^d$. Since

$$F(m, C) = \mathbb{E}[V(m + C\xi)] - \log\det C,$$

differentiation under the expectation gives

$$\nabla_m F(m, C) = \mathbb{E}[\nabla V(X)], \qquad \nabla_C F(m, C) = \mathrm{sym}\,\mathbb{E}[\nabla V(X)\xi^\top] - C^{-1}.$$

Moreover,

$$\nabla_x \log q_{m,C}(X) = -C^{-1}\xi.$$

Thus

$$\mathbb{E}[G^{\mathrm{STL}}(m, C; \xi)] = \nabla F(m, C).$$

Stein's identity gives

$$\mathbb{E}[\nabla V(X)\xi^\top] = \mathbb{E}[\nabla^2 V(X)C],$$

and therefore

$$\mathbb{E}[G^{\text{Price-STL}}(m, C; \xi)] = \nabla F(m, C).$$

The following matrix inequality is useful for controlling the score-estimator variance in symmetric-root coordinates.

Lemma 25. *Let $A \in \mathbb{S}^d$, $C \in \mathbb{S}^d_{++}$, and $B := AC - C^{-1}$. Then*

$$\|B\|_F^2 \le 2\|\operatorname{sym}(B)\|_F^2.$$

Proof. Since A and C are symmetric,

$$CB = CAC - I = B^\top C.$$

Diagonalize $C = Q\operatorname{diag}(c_1, \dots, c_d)Q^\top$ and write $\widetilde{B} := Q^\top BQ$. Then

$$c_i \widetilde{B}_{ij} = c_j \widetilde{B}_{ji}.$$

Let $S := \operatorname{sym}(\widetilde{B})$ and $K := \operatorname{skew}(\widetilde{B})$ where $\operatorname{skew}(M) := \frac{M - M^\top}{2}$. Then

$$K_{ij} = \frac{c_j - c_i}{c_i + c_j} S_{ij}.$$

Since $c_i, c_j > 0$,

$$\|K\|_F^2 \le \|S\|_F^2.$$

By orthogonal invariance and Frobenius orthogonality,

$$\|B\|_F^2 = \|S\|_F^2 + \|K\|_F^2 \le 2\|S\|_F^2 = 2\|\operatorname{sym}(B)\|_F^2.$$

□

We can now establish the explicit relative and additive moment bounds.

Lemma 26. *If V is quadratic, then for every $(m, C) \in \mathbb{R}^d \times \mathbb{S}^d_{++}$,*

$$\mathbb{E}\|G_b^{\mathrm{STL}}(m, C)\|^2 \le \left(1 + \frac{d+4}{b}\right) \|\nabla F(m, C)\|^2,$$

and

$$\mathbb{E}\|G_b^{\text{Price-STL}}(m, C)\|^2 \le \left(1 + \frac{2}{b}\right) \|\nabla F(m, C)\|^2.$$

Under Assumptions 1 and 2, for every $(m, C) \in K_C$,

$$\mathbb{E}\|G_b^{\mathrm{STL}}(m, C)\|^2 \le \left(1 + \frac{4d+16}{b}\right) \|\nabla F(m, C)\|^2 + \frac{128 L_H^2 d(d+2)(d+5)}{\alpha^2 b},$$

and

$$\mathbb{E}\|G_b^{\text{Price-STL}}(m, C)\|^2 \le \left(1 + \frac{8}{b}\right) \|\nabla F(m, C)\|^2 + \frac{16 L_H^2}{\alpha^2 b} \left(10d^2 + \frac{d(d+2)}{2}\right).$$

Proof. Fix (m, C) and write

$$X := m + C\xi, \qquad \xi \sim \mathcal{N}(0, I_d), \qquad g := \nabla F(m, C).$$

For a square-integrable random vector or matrix Z, write

$$\operatorname{Var}(Z) := \mathbb{E}\|Z - \mathbb{E}Z\|^2.$$

If $Z_1, \dots, Z_b$ are independent copies of Z, then

$$\mathbb{E}\left\| \frac{1}{b} \sum_{i=1}^b Z_i \right\|^2 = \|\mathbb{E}Z\|^2 + \frac{1}{b} \operatorname{Var}(Z).$$

Thus it suffices to control the single-sample variances. Define

$$u := \nabla V(m), \qquad H := \nabla^2 V(m), \qquad B := HC - C^{-1}, \qquad S := \operatorname{sym}(B).$$

Suppose first that V is quadratic. Then

$$r(\xi) := \nabla V(X) - C^{-1}\xi = u + B\xi,$$

and

$$G^{\mathrm{STL}} = \big(u + B\xi, \mathrm{sym}(r(\xi)\xi^\top)\big), \qquad G^{\text{Price-STL}} = (u + B\xi, S),$$

with $g = (u, S)$. For STL,

$$\mathbb{E}\|r(\xi) - u\|^2 = \|B\|_F^2.$$

For every $p, z \in \mathbb{R}^d$,

$$\|\,\mathrm{sym}(pz^\top)\|_F^2 = \frac{1}{2}\|p\|^2\|z\|^2 + \frac{1}{2}(p^\top z)^2.$$

Using vanishing odd Gaussian moments,

$$\mathbb{E}[\|r(\xi)\|^2\|\xi\|^2] = d\|u\|^2 + (d+2)\|B\|_F^2.$$

Moreover,

$$\mathbb{E}[(r(\xi)^\top \xi)^2] = \|u\|^2 + 2\|S\|_F^2 + (\mathrm{tr}\, S)^2.$$

Since $\mathbb{E}[\mathrm{sym}(r(\xi)\xi^\top)] = S$,

$$\mathrm{Var}(G^{\mathrm{STL}}) = \frac{d+1}{2}\|u\|^2 + \frac{d+4}{2}\|B\|_F^2 + \frac{1}{2}(\mathrm{tr}\, S)^2.$$

Diagonalize

$$C = Q\,\mathrm{diag}(c_1, \dots, c_d)Q^\top, \qquad \widetilde{B} := Q^\top BQ, \qquad \widetilde{S} := Q^\top SQ.$$

The identity $CB = B^\top C$ gives

$$c_i\widetilde{B}_{ij} = c_j\widetilde{B}_{ji}, \qquad \widetilde{B}_{ij} = \frac{2c_j}{c_i + c_j}\widetilde{S}_{ij}.$$

Hence

$$\|B\|_F^2 \le \sum_i \widetilde{S}_{ii}^2 + 2\sum_{i\neq j}\widetilde{S}_{ij}^2,$$

while

$$(\mathrm{tr}\, S)^2 \le d\sum_i \widetilde{S}_{ii}^2.$$

Substituting into the variance formula,

$$\mathrm{Var}(G^{\mathrm{STL}}) \le (d+4)(\|u\|^2 + \|S\|_F^2) = (d+4)\|g\|^2.$$

For Price–STL, the covariance-factor component is deterministic, so Lemma 25 gives

$$\mathrm{Var}(G^{\text{Price-STL}}) = \|B\|_F^2 \le 2\|S\|_F^2 \le 2\|g\|^2.$$

This proves the quadratic bounds after batch averaging. Now suppose Assumptions 1 and 2 hold and $(m, C) \in K_C$. Then

$$\|C\|_{\mathrm{op}} \le \frac{2}{\sqrt{\alpha}}, \qquad \|C\|_F \le 2\sqrt{\frac{d}{\alpha}}.$$

Define

$$w(\xi) := \nabla V(X) - u - HC\xi, \qquad D(\xi) := \mathrm{sym}((\nabla^2 V(X) - H)C), \qquad M := \frac{2L_H}{\alpha}.$$

Taylor's theorem and Hessian Lipschitzness give

$$\|w(\xi)\| \le \frac{L_H}{2}\|C\xi\|^2 \le M\|\xi\|^2$$

and

$$\|D(\xi)\|_F \le \|\nabla^2 V(X) - H\|_{\mathrm{op}}\|C\|_F \le 2M\sqrt{d}\,\|\xi\|.$$

Therefore

$$\mathbb{E}\|w(\xi)\|^2 \le M^2 d(d+2), \qquad \mathbb{E}\|D(\xi)\|_F^2 \le 4M^2 d^2.$$

Stein's identity gives

$$\mathrm{sym}\,\mathbb{E}[w(\xi)\xi^\top] = \mathbb{E}D(\xi).$$

Define $\delta := (\mathbb{E}w, \mathbb{E}D)$. Then $g = (u, S) + \delta$. Moreover,

$$\|\mathbb{E}w\| \le Md, \qquad \|\mathbb{E}D\|_F \le 2Md,$$

so $\|\delta\|^2 \le 5M^2d^2$. For STL, decompose

$$G^{\mathrm{STL}} = \mathcal{A} + \mathcal{E},$$

where $\mathcal{A}(\xi) := \big(u + B\xi, \mathrm{sym}((u + B\xi)\xi^\top)\big)$ and $\mathcal{E}(\xi) := \big(w(\xi), \mathrm{sym}(w(\xi)\xi^\top)\big)$. Then

$$\mathbb{E}\mathcal{E} = \delta, \qquad \mathbb{E}\mathcal{A} = g - \delta.$$

The quadratic calculation yields

$$\mathrm{Var}(\mathcal{A}) \le (d + 4)\|g - \delta\|^2.$$

Also,

$$\mathbb{E}\|\mathcal{E}\|^2 \le \mathbb{E}[(1 + \|\xi\|^2)\|w(\xi)\|^2] \le M^2 d(d + 2)(d + 5) =: \sigma_H^2.$$

Centering the two components and using $\|a + b\|^2 \le 2\|a\|^2 + 2\|b\|^2$,

$$\mathrm{Var}(G^{\mathrm{STL}}) \le 2(d + 4)\|g - \delta\|^2 + 2\sigma_H^2.$$

Thus

$$\mathrm{Var}(G^{\mathrm{STL}}) \le 4(d + 4)\|g\|^2 + 4(d + 4)\|\delta\|^2 + 2\sigma_H^2.$$

The bound on δ gives

$$(d + 4)\|\delta\|^2 \le \frac{15}{2}\sigma_H^2.$$

Hence

$$\mathrm{Var}(G^{\mathrm{STL}}) \le (4d + 16)\|g\|^2 + 32\sigma_H^2$$

and therefore

$$\mathrm{Var}(G^{\mathrm{STL}}) \le (4d + 16)\|g\|^2 + \frac{128L_H^2 d(d + 2)(d + 5)}{\alpha^2}.$$

For Price–STL,

$$G^{\text{Price-STL}} = (u + B\xi + w(\xi), S + D(\xi)),$$

so

$$G^{\text{Price-STL}} - g = (B\xi, 0) + (w - \mathbb{E}w, D - \mathbb{E}D).$$

Therefore

$$\mathrm{Var}(G^{\text{Price-STL}}) \le 2\|B\|_F^2 + 2\mathbb{E}\|w\|^2 + 2\mathbb{E}\|D\|_F^2.$$

Let $g_C := \nabla_C F(m, C)$. Since $g_C = S + \mathbb{E}D$, Lemma 25 and Jensen's inequality give

$$\|B\|_F^2 \le 2\|S\|_F^2 \le 4\|g_C\|_F^2 + 4\|\mathbb{E}D\|_F^2 \le 4\|g\|^2 + 4\mathbb{E}\|D\|_F^2.$$

Hence

$$\mathrm{Var}(G^{\text{Price-STL}}) \le 8\|g\|^2 + 2\mathbb{E}\|w\|^2 + 10\mathbb{E}\|D\|_F^2,$$

which yields

$$\mathrm{Var}(G^{\text{Price-STL}}) \le 8\|g\|^2 + \frac{16L_H^2}{\alpha^2}\left(10d^2 + \frac{d(d + 2)}{2}\right).$$

Batch averaging completes the proof. □

We next construct the smooth spectral extension needed to apply projected acceleration.

Using the scalar function ψ from equation 2, define

$$\widetilde{h}(C) := \sum_{i=1}^{d} \psi(\lambda_i(C)), \qquad \widetilde{F}(m, C) := E(m, C) + \widetilde{h}(C).$$

Lemma 27. *Under Assumption 1, the minimizer $x_\star = (m_\star, C_\star)$ belongs to the interior of K_C. Moreover, $\widetilde{F}$ is globally α-strongly convex and 5β-smooth on $\mathbb{R}^d \times \mathbb{S}^d$, satisfies*

$$\widetilde{F} = F, \qquad \nabla\widetilde{F} = \nabla F \qquad \text{on } K_C,$$

and obeys

$$\widetilde{F}(P_{K_C} z) \le \widetilde{F}(z)$$

for every $z \in \mathbb{R}^d \times \mathbb{S}^d$.

Proof. From $\beta^{-1}I \preceq \Sigma_\star \preceq \alpha^{-1}I$ and $C_\star = \Sigma_\star^{1/2}$,

$$\frac{1}{\sqrt{\beta}}I \preceq C_\star \preceq \frac{1}{\sqrt{\alpha}}I,$$

so $x_\star$ belongs to the interior of K_C.

Since $0 \le \psi'' \le 4\beta$, the spectral function $\widetilde{h}$ is convex and 4β-smooth on $\mathbb{S}^d$. By Lemma 1, E is α-strongly convex and β-smooth. Hence $\widetilde{F}$ is globally α-strongly convex and 5β-smooth. For every $(m, C) \in K_C$,

$$\widetilde{h}(C) = -\sum_{i=1}^{d} \log \lambda_i(C) = -\log\det C.$$

The two branches of ψ have matching first derivatives, so the gradients also agree.

It remains to prove projection monotonicity. Write

$$C = Q\,\mathrm{diag}(c_1, \dots, c_d)Q^\top$$

and let $\widetilde{c}_i$ be c_i clipped to $[1/(2\sqrt{\beta}), 2/\sqrt{\alpha}]$. Then

$$P_{K_C}(m, C) = \left(m, Q\,\mathrm{diag}(\widetilde{c}_1, \dots, \widetilde{c}_d)Q^\top\right).$$

Let q_i be the ith column of Q and define

$$X_c := m + \sum_{j=1}^{d} c_j q_j \xi_j.$$

Gaussian integration by parts gives

$$\frac{\partial}{\partial c_i}\mathbb{E}[V(X_c)] = c_i\,\mathbb{E}[q_i^\top \nabla^2 V(X_c) q_i].$$

Write $a_i(c) := \mathbb{E}[q_i^\top \nabla^2 V(X_c) q_i]$, so that $\alpha \le a_i(c) \le \beta$. The derivative of $\widetilde{F}$ in the c_i coordinate is $c_i a_i(c) + \psi'(c_i)$. If $c_i \ge 2/\sqrt{\alpha}$, then $\psi'(c_i) = -1/c_i$, and

$$c_i a_i(c) + \psi'(c_i) \ge \alpha c_i - \frac{1}{c_i} \ge 2\sqrt{\alpha} - \frac{\sqrt{\alpha}}{2} > 0.$$

Thus clipping c_i downward cannot increase $\widetilde{F}$. If $c_i \le 1/(2\sqrt{\beta})$, then

$$\psi'(c_i) = 4\beta c_i - 4\sqrt{\beta},$$

and hence

$$c_i a_i(c) + \psi'(c_i) = c_i(a_i(c) + 4\beta) - 4\sqrt{\beta} \le \frac{5\beta}{2\sqrt{\beta}} - 4\sqrt{\beta} < 0.$$

Thus clipping c_i upward cannot increase $\widetilde{F}$. Coordinates already in the spectral interval are unchanged, proving $\widetilde{F}(P_{K_C}(m, C)) \le \widetilde{F}(m, C)$. □

The remaining optimization argument can be done by the following projected acceleration lemma, which extends the unconstrained strong-growth acceleration argument of Vaswani et al. (2019) to the projected setting.

Lemma 28. *Let K be a nonempty closed convex subset of a finite-dimensional Euclidean space. Let Φ be globally α-strongly convex and β-smooth, with minimizer $x_\star \in K$, and write $\Phi^\star := \Phi(x_\star)$ and $\kappa := \beta/\alpha$. Suppose*

$$\Phi(P_K z) \le \Phi(z)$$

for every z in the ambient space. Set

$$s := \sqrt{\max\{3\kappa, 4\}}, \qquad \eta := \frac{1}{\alpha s^2}.$$

Starting from $x_0 = v_0 \in K$, consider the updates

$$y_k := \frac{x_k + s^{-1}v_k}{1 + s^{-1}},$$

$$x_{k+1} := P_K(y_k - \eta g_k),$$

and

$$v_{k+1} := P_K\left((1 - s^{-1})v_k + s^{-1}y_k - \frac{1}{\alpha s}g_k\right).$$

Suppose, conditionally on the past $\mathcal{F}_k$,

$$\mathbb{E}[g_k \mid \mathcal{F}_k] = \nabla\Phi(y_k)$$

and

$$\mathbb{E}[\|g_k - \nabla\Phi(y_k)\|^2 \mid \mathcal{F}_k] \le \frac{1}{2}\|\nabla\Phi(y_k)\|^2 + \sigma^2.$$

Then, for every integer $T \ge 0$,

$$\mathbb{E}[\Phi(x_T) - \Phi^\star] \le e^{-T/s}\left(\Phi(x_0) - \Phi^\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2\right) + \frac{\sigma^2}{\alpha s}.$$

Proof. Since $s \ge 2$ and $s^2 \ge 3\kappa = 3\beta/\alpha$, we have $\beta\eta \le 1/3$. The projection steps ensure that $x_k, v_k \in K$, and convexity of K gives $y_k \in K$. Let $\mathcal{F}_k$ denote the sigma-field generated by the randomness before the fresh batch at iteration k, so that x_k, v_k, and y_k are $\mathcal{F}_k$-measurable.

Conditional unbiasedness and the variance assumption imply

$$\mathbb{E}[\|g_k\|^2 \mid \mathcal{F}_k] = \|\nabla\Phi(y_k)\|^2 + \mathbb{E}[\|g_k - \nabla\Phi(y_k)\|^2 \mid \mathcal{F}_k] \le \frac{3}{2}\|\nabla\Phi(y_k)\|^2 + \sigma^2.$$

Projection optimality for the v-update gives

$$p_k := \alpha s\left((1 - s^{-1})v_k + s^{-1}y_k - v_{k+1}\right) - g_k$$

with

$$\langle p_k, z - v_{k+1}\rangle \le 0 \qquad \text{for every } z \in K.$$

Define

$$\Lambda_k(z) := \Phi(y_k) + \langle g_k, z - y_k\rangle + \frac{\alpha}{2}\|z - y_k\|^2 + \langle p_k, z - v_{k+1}\rangle.$$

Thus strong convexity and conditional unbiasedness give

$$\mathbb{E}[\Lambda_k(z) \mid \mathcal{F}_k] \le \Phi(y_k) + \langle \nabla\Phi(y_k), z - y_k\rangle + \frac{\alpha}{2}\|z - y_k\|^2 \le \Phi(z)$$

for every $\mathcal{F}_k$-measurable $z \in K$.

Projection monotonicity and β-smoothness give

$$\Phi(x_{k+1}) \le \Phi(y_k - \eta g_k) \le \Phi(y_k) - \eta\langle \nabla\Phi(y_k), g_k\rangle + \frac{\beta\eta^2}{2}\|g_k\|^2.$$

Taking conditional expectations,

$$\mathbb{E}[\Phi(x_{k+1}) \mid \mathcal{F}_k] \le \Phi(y_k) - \eta\|\nabla\Phi(y_k)\|^2 + \frac{\beta\eta^2}{2}\mathbb{E}[\|g_k\|^2 \mid \mathcal{F}_k].$$

Define

$$\varphi_0(z) := \Phi(x_0) + \frac{\alpha}{2}\|z - x_0\|^2, \qquad \varphi_{k+1} := (1 - s^{-1})\varphi_k + s^{-1}\Lambda_k.$$

Every φ_k is quadratic with Hessian αI. Its unique minimizer is v_k. Indeed, this holds at $k = 0$, and

$$\nabla\varphi_{k+1}(v_{k+1}) = (1 - s^{-1})\alpha(v_{k+1} - v_k) + s^{-1}\left(g_k + p_k + \alpha(v_{k+1} - y_k)\right) = 0$$

by the definition of p_k. Writing

$$\varphi_k^{\min} := \varphi_k(v_k),$$

we have

$$\varphi_k(z) = \varphi_k^{\min} + \frac{\alpha}{2}\|z - v_k\|^2.$$

The definition of y_k gives

$$x_k - y_k = -s^{-1}(v_k - y_k).$$

Set

$$\zeta_0 := 0, \qquad \zeta_{k+1} := (1 - s^{-1})\zeta_k + \eta\sigma^2.$$

We prove by induction that

$$\mathbb{E}[\varphi_k^{\min}] \geq \mathbb{E}[\Phi(x_k)] - \zeta_k.$$

The claim holds with equality at $k = 0$.

Evaluating φ_{k+1} at v_{k+1} gives

$$\varphi_{k+1}^{\min} = (1 - s^{-1})\varphi_k^{\min} + s^{-1}\Phi(y_k) + (1 - s^{-1})\frac{\alpha}{2}\|v_{k+1} - v_k\|^2 + s^{-1}\langle g_k, v_{k+1} - y_k\rangle + s^{-1}\frac{\alpha}{2}\|v_{k+1} - y_k\|^2.$$

Completing the square in the terms involving v_{k+1} gives the exact identity

$$\begin{aligned}&(1 - s^{-1})\frac{\alpha}{2}\|v_{k+1} - v_k\|^2 + s^{-1}\langle g_k, v_{k+1} - y_k\rangle + s^{-1}\frac{\alpha}{2}\|v_{k+1} - y_k\|^2 \\ &= (1 - s^{-1})s^{-1}\left(\langle g_k, v_k - y_k\rangle + \frac{\alpha}{2}\|v_k - y_k\|^2\right) \\ &- \frac{1}{2\alpha s^2}\|g_k\|^2 + \frac{\alpha}{2}\left\|v_{k+1} - \left((1 - s^{-1})v_k + s^{-1}y_k - \frac{1}{\alpha s}g_k\right)\right\|^2.\end{aligned}$$

Dropping the final nonnegative term,

$$\varphi_{k+1}^{\min} \geq (1 - s^{-1})\varphi_k^{\min} + s^{-1}\Phi(y_k) + (1 - s^{-1})s^{-1}\left(\langle g_k, v_k - y_k\rangle + \frac{\alpha}{2}\|v_k - y_k\|^2\right) - \frac{\eta}{2}\|g_k\|^2.$$

By the induction hypothesis, $\mathbb{E}[\varphi_k^{\min}] \geq \mathbb{E}[\Phi(x_k)] - \zeta_k$. Strong convexity gives

$$\Phi(x_k) \geq \Phi(y_k) + \langle \nabla\Phi(y_k), x_k - y_k\rangle + \frac{\alpha}{2}\|x_k - y_k\|^2.$$

Using $x_k - y_k = -s^{-1}(v_k - y_k)$ and conditional unbiasedness,

$$\mathbb{E}[\Phi(x_k)] \geq \mathbb{E}[\Phi(y_k)] - s^{-1}\mathbb{E}\langle g_k, v_k - y_k\rangle + \frac{\alpha}{2s^2}\mathbb{E}\|v_k - y_k\|^2.$$

Substituting into the estimate for $\varphi_{k+1}^{\min}$ cancels the inner-product terms and yields

$$\mathbb{E}[\varphi_{k+1}^{\min}] \geq \mathbb{E}[\Phi(y_k)] - (1 - s^{-1})\zeta_k - \frac{\eta}{2}\mathbb{E}\|g_k\|^2 + \frac{\alpha}{2}(1 - s^{-1})s^{-1}(1 + s^{-1})\mathbb{E}\|v_k - y_k\|^2.$$

Dropping the final nonnegative term,

$$\mathbb{E}[\varphi_{k+1}^{\min}] \geq \mathbb{E}[\Phi(y_k)] - (1 - s^{-1})\zeta_k - \frac{\eta}{2}\mathbb{E}\|g_k\|^2.$$

The one-step upper bound implies

$$\mathbb{E}[\Phi(y_k)] \geq \mathbb{E}[\Phi(x_{k+1})] + \eta\mathbb{E}\|\nabla\Phi(y_k)\|^2 - \frac{\beta\eta^2}{2}\mathbb{E}\|g_k\|^2.$$

Therefore

$$\mathbb{E}[\varphi_{k+1}^{\min}] \geq \mathbb{E}[\Phi(x_{k+1})] - (1 - s^{-1})\zeta_k + \eta\mathbb{E}\|\nabla\Phi(y_k)\|^2 - \frac{\eta}{2}(1 + \beta\eta)\mathbb{E}\|g_k\|^2.$$

Since $\beta\eta \leq 1/3$,

$$\eta\mathbb{E}\|\nabla\Phi(y_k)\|^2 - \frac{\eta}{2}(1 + \beta\eta)\mathbb{E}\|g_k\|^2 \geq -\eta\sigma^2.$$

Hence

$$\mathbb{E}[\varphi_{k+1}^{\min}] \geq \mathbb{E}[\Phi(x_{k+1})] - \zeta_{k+1},$$

closing the induction. Finally, since $x_\star \in K$,

$$\mathbb{E}[\Lambda_k(x_\star) \mid \mathcal{F}_k] \leq \Phi^\star.$$

Iterating the estimate-sequence recursion gives

$$\mathbb{E}[\varphi_T(x_\star)] - \Phi^\star \leq (1 - s^{-1})^T\left(\varphi_0(x_\star) - \Phi^\star\right).$$

The induction invariant and $\varphi_T^{\min} \le \varphi_T(x_\star)$ imply

$$\mathbb{E}[\Phi(x_T) - \Phi^\star] \le (1 - s^{-1})^T \left(\Phi(x_0) - \Phi^\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2\right) + \zeta_T.$$

Moreover,

$$\zeta_T = \eta\sigma^2 \sum_{j=0}^{T-1} (1 - s^{-1})^j \le \eta s \sigma^2 = \frac{\sigma^2}{\alpha s}.$$

Using $(1 - s^{-1})^T \le e^{-T/s}$ completes the proof. □

Combining the moment bounds, the smooth extension, and the projected acceleration lemma gives the full score-based guarantee.

Theorem 29 (Full version of Theorem 5)**.** *Suppose Assumptions 1 and 2 hold, $x_0 \in K_C$, and $0 < \varepsilon \le \Delta_0$, where*

$$\Delta_0 := F(q_0) - F^\star + \frac{\alpha}{2}\|x_0 - x_\star\|^2.$$

Let

$$s := \sqrt{15\kappa}, \qquad T_\varepsilon := \left\lceil s \log \frac{2\Delta_0}{\varepsilon} \right\rceil.$$

Choose

$$\eta := \frac{1}{\alpha s^2} = \frac{1}{15\beta}.$$

For STL, choose

$$b_{\mathrm{STL}} := \left\lceil \max\left\{8d + 32, \frac{256\lambda_H^2 d(d+2)(d+5)}{s\varepsilon}\right\} \right\rceil.$$

For Price–STL, choose

$$b_{\mathrm{Price}} := \left\lceil \max\left\{16, \frac{16\lambda_H^2 d(21d+2)}{s\varepsilon}\right\} \right\rceil.$$

Then Algorithm 2, with either estimator and its corresponding fixed batch size, satisfies

$$\mathbb{E}[F(q_{T_\varepsilon}) - F^\star] \le \varepsilon.$$

The exact oracle counts are

$$N_{\nabla V}^{\mathrm{STL}} = b_{\mathrm{STL}} T_\varepsilon,$$

and

$$N_{\nabla V}^{\mathrm{Price}} = b_{\mathrm{Price}} T_\varepsilon, \qquad N_{\mathrm{HVP}}^{\mathrm{Price}} = d b_{\mathrm{Price}} T_\varepsilon.$$

Consequently,

$$T_\varepsilon = O\left(\sqrt{\kappa} \log \frac{2\Delta_0}{\varepsilon}\right),$$

$$N_{\nabla V}^{\mathrm{STL}} = O\left(\left(d\sqrt{\kappa} + \frac{\lambda_H^2 d^3}{\varepsilon}\right) \log \frac{2\Delta_0}{\varepsilon}\right),$$

$$N_{\nabla V}^{\mathrm{Price}} = O\left(\left(\sqrt{\kappa} + \frac{\lambda_H^2 d^2}{\varepsilon}\right) \log \frac{2\Delta_0}{\varepsilon}\right),$$

and

$$N_{\mathrm{HVP}}^{\mathrm{Price}} = O\left(\left(d\sqrt{\kappa} + \frac{\lambda_H^2 d^3}{\varepsilon}\right) \log \frac{2\Delta_0}{\varepsilon}\right).$$

Proof. The two projection steps ensure that $x_k, v_k \in K_C$, and convexity gives $y_k \in K_C$. By Lemma 27, $\widetilde{F}$ is globally α-strongly convex and 5β-smooth, agrees with F and its gradient on K_C, and satisfies

$$\widetilde{F}(P_{K_C} z) \le \widetilde{F}(z).$$

Since $x_\star$ lies in the interior of K_C and $\nabla F(x_\star) = 0$, it is also the global minimizer of $\widetilde{F}$.

For the selected estimator, define

$$\sigma^2 := \begin{cases} \dfrac{128L_H^2 d(d+2)(d+5)}{\alpha^2}, & \text{STL}, \\ \dfrac{8L_H^2 d(21d+2)}{\alpha^2}, & \text{Price–STL}. \end{cases}$$

Lemma 26 and fresh batch sampling give

$$\mathbb{E}[g_k \mid \mathcal{F}_k] = \nabla \widetilde{F}(y_k).$$

For STL,

$$\mathbb{E}[\|g_k - \nabla \widetilde{F}(y_k)\|^2 \mid \mathcal{F}_k] \le \frac{4d+16}{b}\|\nabla \widetilde{F}(y_k)\|^2 + \frac{\sigma^2}{b},$$

while for Price–STL,

$$\mathbb{E}[\|g_k - \nabla \widetilde{F}(y_k)\|^2 \mid \mathcal{F}_k] \le \frac{8}{b}\|\nabla \widetilde{F}(y_k)\|^2 + \frac{\sigma^2}{b}.$$

The prescribed batch sizes satisfy $b_{\mathrm{STL}} \ge 8d + 32$ and $b_{\mathrm{Price}} \ge 16$, so in either case

$$\mathbb{E}[\|g_k - \nabla \widetilde{F}(y_k)\|^2 \mid \mathcal{F}_k] \le \frac{1}{2}\|\nabla \widetilde{F}(y_k)\|^2 + \frac{\sigma^2}{b}.$$

Lemma 28 therefore applies with $\Phi = \widetilde{F}$, $K = K_C$, strong-convexity constant α, smoothness constant 5β, and additive variance parameter σ^2/b. Its parameter becomes

$$s = \sqrt{\max\{15\kappa, 4\}} = \sqrt{15\kappa}.$$

Hence

$$\mathbb{E}[F(q_T) - F^\star] \le e^{-T/s}\Delta_0 + \frac{\sigma^2}{\alpha b s}.$$

At $T = T_\varepsilon$, $e^{-T_\varepsilon/s}\Delta_0 \le \frac{\varepsilon}{2}$. Using $\lambda_H^2 = L_H^2/\alpha^3$, each prescribed batch size also satisfies

$$b \ge \frac{2\sigma^2}{\alpha s \varepsilon},$$

and therefore

$$\frac{\sigma^2}{\alpha b s} \le \frac{\varepsilon}{2}.$$

This proves the accuracy bound.

Each STL sample uses one target-gradient evaluation. Each Price–STL sample uses one target-gradient evaluation and d HVPs to form $\nabla^2 V(X)C$. The exact oracle counts and their asymptotic forms follow directly from the batch sizes and T_ε. □

For quadratic targets, $L_H = 0$, so the additive variance term vanishes. The preceding proof gives geometric accelerated convergence with fixed batch sizes of orders d and 1 for STL and Price–STL, respectively.

Remark 3. The fixed-batch choices in Theorems 4 and 5 are selected for the target accuracy ε and kept fixed throughout the run. Consequently, the variance-dependent oracle terms carry an additional factor $\log(2\Delta_0/\varepsilon)$. Multistage or increasing-batch stochastic acceleration can avoid this extra logarithmic factor (Ghadimi & Lan, 2013).

C.4 Dimension dependence and structured Hessian variation

The additive variance bounds underlying Table 1 scale as $L_H^2 d^3/\alpha^2$ for FD and STL, and as $L_H^2 d^2/\alpha^2$ for their second-order variants. These powers arise from uniform norm and Gaussian-moment bounds. Here we give a sharper analysis of the same Price–STL estimator using two alternative measures of Hessian variation. Write $H(x) := \nabla^2 V(x)$ and define the global curvature oscillation

$$\omega_H := \inf_{A \in \mathbb{S}^d} \sup_{x \in \mathbb{R}^d} \|H(x) - A\|_{\mathrm{op}}.$$

Under Assumption 1, $\omega_H \le (\beta-\alpha)/2$. Alternatively, suppose a constant $L_{H,F}$ satisfies

$$\|H(x)-H(y)\|_F \le L_{H,F}\|x-y\|, \qquad x,y\in\mathbb{R}^d. \tag{3}$$

These quantities measure the amplitude and the rate of Hessian variation, respectively. For example, $V''(x)=a+\delta\sin(kx)$, with $a>\delta>0$ and $k>0$, has $\omega_H=\delta$ and $L_{H,F}=\delta k$ while its curvature bounds stay fixed.

Proposition 30. *Under Assumption 1, the single-sample Price–STL estimator satisfies, for every $(m,C)\in K_C$,*

$$\mathrm{Var}(G^{\text{Price-STL}}) \le 2\|\nabla F(m,C)\|^2+\sigma_{\text{str}}^2, \qquad \sigma_{\text{str}}^2 := \frac{8\omega_H^2 d}{\alpha}.$$

If equation 3 also holds, one may instead take

$$\sigma_{\text{str}}^2 := \min\left\{\frac{8\omega_H^2 d}{\alpha}, \frac{32L_{H,F}^2 d}{\alpha^2}\right\}.$$

Proof. Let $X=m+C\xi$, $\bar H=\mathbb{E}H(X)$, $u=\mathbb{E}\nabla V(X)$, and $B=\bar H C-C^{-1}$. Stein's identity gives

$$e := \nabla V(X)-u-\bar H C\xi, \qquad \mathbb{E}e=0, \qquad \mathbb{E}[e\xi^\top]=0.$$

Consequently, the mean component has variance $\|B\|_F^2+\mathbb{E}\|e\|^2$. Gaussian Poincaré gives

$$\mathbb{E}\|e\|^2 \le D_C := \mathbb{E}\|(H(X)-\bar H)C\|_F^2.$$

The covariance-factor component $\mathrm{sym}(H(X)C)-C^{-1}$ also has variance at most D_C. Since $CB=B^\top C$ and $\nabla_C F=\mathrm{sym}(B)$, Lemma 25 yields

$$\mathrm{Var}(G^{\text{Price-STL}}) \le 2\|\nabla_C F\|_F^2+2D_C.$$

For any constant symmetric A, centering minimizes the squared deviation, so

$$D_C \le \mathbb{E}\|(H(X)-A)C\|_F^2 \le \sup_x \|H(x)-A\|_{\text{op}}^2\|C\|_F^2.$$

Taking the infimum gives $D_C\le\omega_H^2\|C\|_F^2$. Under equation 3, an independent copy Y of X gives the alternative bound

$$D_C=\tfrac12\mathbb{E}\|(H(X)-H(Y))C\|_F^2 \le \tfrac12 L_{H,F}^2\|C\|_{\text{op}}^2\mathbb{E}\|X-Y\|^2 = L_{H,F}^2\|C\|_{\text{op}}^2\|C\|_F^2.$$

Finally, $\|C\|_{\text{op}}^2\le 4/\alpha$ and $\|C\|_F^2\le 4d/\alpha$ on K_C give the stated constants. □

Fixed-batch consequence. Let $x_0\in K_C$, $\Delta_0=F(q_0)-F^\star+\frac{\alpha}{2}\|x_0-x_\star\|^2$, and $0<\varepsilon\le\Delta_0$. Keep Algorithm 2 and set

$$s=\sqrt{15\kappa}, \quad \eta=\frac{1}{15\beta}, \quad b=\left\lceil \max\left\{16, \frac{2\sigma_{\text{str}}^2}{\alpha s\varepsilon}\right\}\right\rceil, \quad T_\varepsilon=\left\lceil s\log\frac{2\Delta_0}{\varepsilon}\right\rceil.$$

The batch variance is at most $2\|\nabla F\|^2/b+\sigma_{\text{str}}^2/b$. Lemmas 27 and 28 therefore give

$$\mathbb{E}[F(q_{T_\varepsilon})-F^\star] \le e^{-T_\varepsilon/s}\Delta_0+\frac{\sigma_{\text{str}}^2}{\alpha b s}\le\varepsilon.$$

Thus the curvature-oscillation bound, without Hessian Lipschitzness, yields

$$N_{\nabla V}=O\left(\left[\sqrt{\kappa}+\frac{\omega_H^2 d}{\alpha^2\varepsilon}\right]\log\frac{2\Delta_0}{\varepsilon}\right), \qquad N_{\text{HVP}}=dN_{\nabla V}. \tag{4}$$

If equation 3 holds, ω_H^2/α^2 in this bound can be replaced by $\min\{\omega_H^2/\alpha^2, L_{H,F}^2/\alpha^3\}$. The sampling term is linear in d for gradient queries, while HVP accounting retains the factor d.

Graph-probit example. For the targets in Appendix D.2,

$$H(f) = Q_\tau + \operatorname{diag}\big(\ell''(\widetilde{y}_i f_i)\big), \qquad 0 < \ell'' < 1.$$

Choosing $A = Q_\tau + \frac{1}{2}I$ gives $\omega_H \le \frac{1}{2}$. Moreover,

$$\|H(f) - H(g)\|_F^2 = \sum_i |\ell''(\widetilde{y}_i f_i) - \ell''(\widetilde{y}_i g_i)|^2 \le C_\Phi^2 \|f - g\|^2,$$

where $C_\Phi := \sup_t |\ell'''(t)| < \infty$, so $L_{H,F} \le C_\Phi$. Both constants are independent of d and τ. Since $\alpha = 1$, the batch choice above gives

$$N_{\nabla V} = O\left(\left[\sqrt{\kappa} + \frac{d}{\varepsilon}\right] \log \frac{2\Delta_0}{\varepsilon}\right), \qquad N_{\mathrm{HVP}} = dN_{\nabla V}$$

for Price–STL. Thus the gradient-query sampling term is linear in d and independent of τ.

D Experimental details

D.1 Gaussian condition-number experiment

For $d = 8$ and $\kappa \in \{10, 30, 100, 300, 1000, 3000\}$, we use

$$\mu = \frac{\mathbf{1}_d}{\sqrt{d}}, \qquad H_\kappa = \frac{1}{2}\operatorname{diag}\Big(\kappa^{j/(d-1)}\Big)_{j=0}^{d-1}, \qquad \pi_\kappa = \mathcal{N}(\mu, H_\kappa^{-1}).$$

Thus $\alpha = 1/2$, $\beta = \kappa/2$, and $q_\star = \pi_\kappa$. We optimize both mean and covariance from $q_0 = \mathcal{N}(0, I_d)$.

At every update, we evaluate the exact gap

$$F(q) - F^\star = \frac{1}{2}\left((m-\mu)^\top H_\kappa (m-\mu) + \operatorname{tr}(H_\kappa \Sigma) - \log\det(H_\kappa \Sigma) - d\right).$$

Stochastic STL scaling. Figure 2 uses Algorithm 2 with G_b^{STL}, $b = 64$, $\eta = 1/(15\beta)$, and $s = \sqrt{15\kappa}$. Its nonaccelerated counterpart uses the same estimator, step, and domain K_C, with the projected-gradient update specified in Appendix D.3. We use the five seeds in Appendix D.4, a budget of 120κ updates, and no restarts or early stopping. The scaling panel records the first update at which the five-seed mean gap is at most 10^{-6}; every run reaches this threshold. STL remains stochastic away from the optimum. On these quadratic targets, Lemma 26 bounds its variance by $((d+4)/b)\|\nabla F\|^2$, with $(d+4)/b = 3/16 < 1/2$, so Lemma 28 applies without an additive variance term.

Paired benchmark. We use the paired estimator in Algorithm 1, with $\eta = 2/\kappa$ and $\rho = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$, and its nonaccelerated variant with $\rho = 0$. The constant Hessian makes each paired sample exact:

$$G^{\mathrm{paired}}(m, L; \xi, \zeta) = \big(H_\kappa(m - \mu), \mathrm{low}(H_\kappa L)\big).$$

We therefore use $b = 1$ for Figure 4. The factor constraint $\|L\|_F \le \sqrt{d/\alpha} = \sqrt{2d}$ contains the initial and optimal factors and was inactive throughout. There are no restarts. We record the first update with gap at most 10^{-6}, with a budget of $\lceil 20\kappa \rceil$ updates; all runs reach this threshold.

Additional methods at $\kappa = 300$. Figure 5 compares all seventeen methods in Table 3 at $b = 1, 8, 64$, with 6000 updates and no early stopping. Steps and domains use $\alpha = 1/2$ and $\beta = 150$, with FD offset $\nu = 10^{-2}/\sqrt{\alpha}$. Gaps are evaluated exactly at every update. The eight FD and paired variants are samplewise exact on this target and use one seed per batch; the other nine methods use the five seeds specified in Appendix D.4.

D.2 OpenML datasets and graph-probit posteriors

We use version 1 of the five OpenML datasets in Table 2, retaining every observation and label. Predictors are centered and divided by their population standard deviation after removing zero-variance columns.

Numerical zeros in Diabetes are retained without imputation. Classes are encoded as $y_i \in \{0, 1\}$, with $\widetilde{y}_i = 2y_i - 1$. Let $X_0 \in \mathbb{R}^{n \times p}$ denote the standardized predictors.

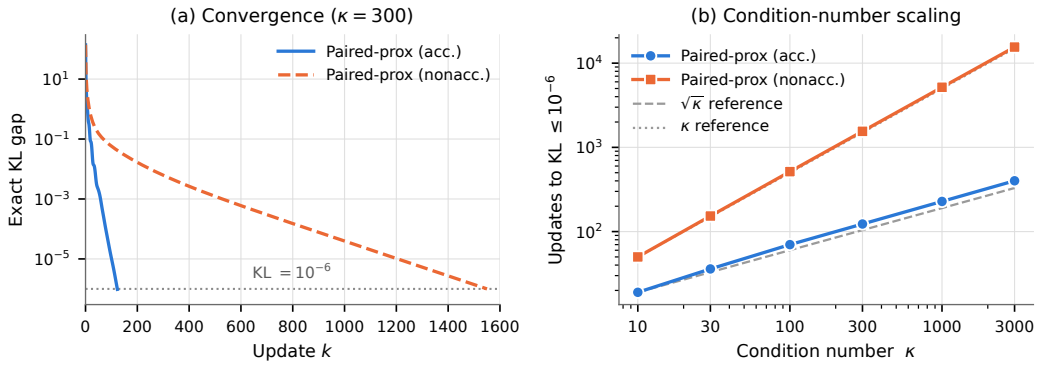


Figure 4: **Paired benchmark on Gaussian targets.** Accelerated and nonaccelerated Paired-prox with a samplewise-exact estimator: (a) KL gap at $\kappa = 300$; (b) updates to KL $\leq 10^{-6}$, with $\sqrt{\kappa}$ and κ reference lines.

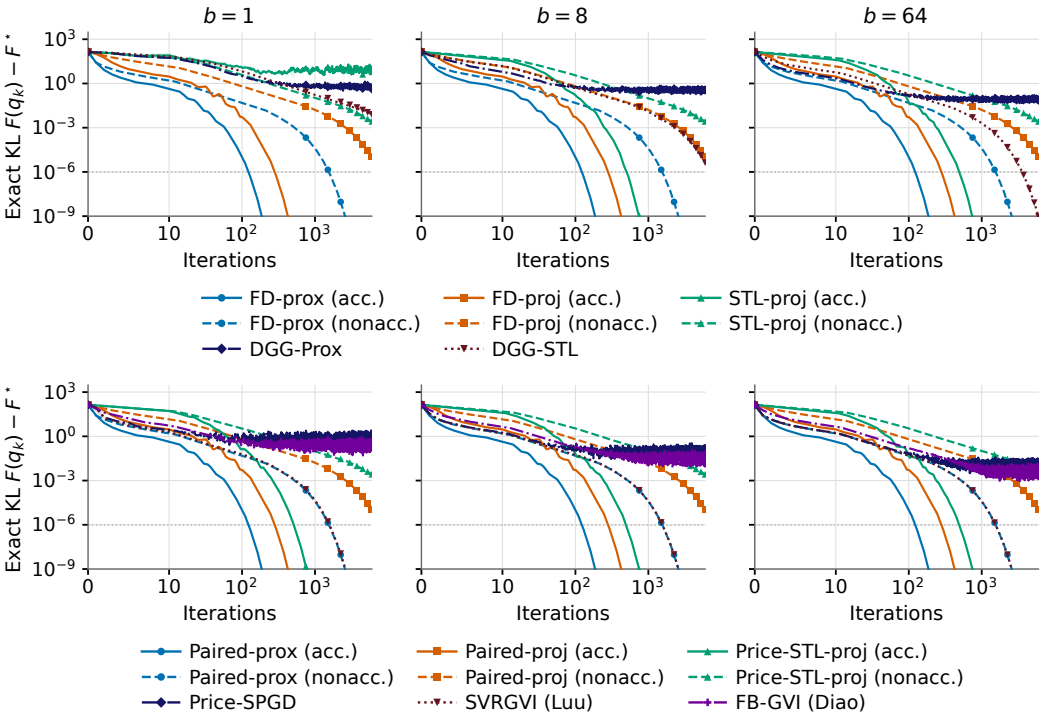


Figure 5: **Gaussian comparisons at $\kappa = 300$.** Rows show first- and second-order methods; columns show $b = 1, 8, 64$. Curves are exact KL gaps, averaged over five seeds with 20–80 percentile bands for stochastic methods. FD and paired methods use one seed. Solid and dashed lines distinguish accelerated and nonaccelerated variants. The dotted line marks 10^{-6}; gaps below 10^{-9} leave the displayed range.

Dataset	OpenML ID	$d = n$	κ
Sonar	40	208	69.47
Ionosphere	59	351	75.64
WDBC	1510	569	69.68
Diabetes	37	768	70.20
Banknote	1462	1372	69.43

Table 2: **Graph-probit dimensions and conditioning.** Each observation has one latent score. All targets use $\alpha = 1$, so $\kappa = \beta$.

Graph-probit classification. We construct an unweighted graph from X_0 using Euclidean ten-nearest neighbors, without self edges and with ties resolved by row index. An edge is included if either endpoint selects the other. For symmetric adjacency matrix W and degree matrix D_G, define

$$L_G = I_n - D_G^{-1/2} W D_G^{-1/2}, \qquad Q_\tau = I_n + \tau L_G.$$

The posterior has one latent score per observation and unit probit noise:

$$V_{\mathrm{probit}}(f) = \frac{1}{2} f^\top Q_\tau f + \sum_{i=1}^{n} \ell(\widetilde{y}_i f_i), \qquad \ell(t) = -\log \Phi(t).$$

This shifted graph-Laplacian prior follows the model family of Miller et al. (2020); see also Bertozzi et al. (2018). We fix $\tau = 49$, giving $\kappa \le 100$.

Writing $r = \phi/\Phi$, where ϕ is the standard normal density, gives

$$\begin{aligned} \ell'(t) &= -r(t), & \ell''(t) &= r(t)(t + r(t)), \\ \ell'''(t) &= r(t)\big[1 - (t + r(t))(t + 2r(t))\big]. \end{aligned} \tag{5}$$

For $Z \sim \mathcal{N}(0,1)$, the truncated-normal identities $\mathbb{E}[Z \mid Z \le t] = -r(t) < t$ and $\mathrm{Var}(Z \mid Z \le t) = 1 - \ell''(t) > 0$ imply $0 < \ell'' < 1$. Hence

$$\begin{aligned} \nabla V_{\mathrm{probit}}(f) &= Q_\tau f - \big(\widetilde{y}_i r(\widetilde{y}_i f_i)\big)_{i=1}^n, \\ \nabla^2 V_{\mathrm{probit}}(f) &= Q_\tau + \mathrm{diag}\big(\ell''(\widetilde{y}_i f_i)\big), \end{aligned}$$

and $I_n \preceq \nabla^2 V_{\mathrm{probit}} \preceq (2 + \tau\lambda_{\max}(L_G)) I_n$. We use $\alpha = 1$ and $\beta = \kappa = 2 + \tau\lambda_{\max}(L_G)$.

The function ℓ''' is continuous and tends to zero at both tails. At $+\infty$ this follows directly from $r = \phi/\Phi$; at $-\infty$, the Gaussian tail expansion gives

$$\frac{\Phi(-x)}{\phi(x)} = x^{-1} - x^{-3} + 3x^{-5} - 15x^{-7} + O(x^{-9}), \quad r(-x) = x + x^{-1} - 2x^{-3} + 10x^{-5} + O(x^{-7}),$$

so $\ell'''(-x) = -2x^{-3} + O(x^{-5})$. Thus $C_\Phi := \sup_t |\ell'''(t)| < \infty$, and

$$\|\nabla^2 V_{\mathrm{probit}}(f) - \nabla^2 V_{\mathrm{probit}}(g)\|_{\mathrm{op}} = \max_i |\ell''(\widetilde{y}_i f_i) - \ell''(\widetilde{y}_i g_i)| \le C_\Phi \|f - g\|.$$

Varying a single coordinate shows that $L_H = C_\Phi$, independently of τ, the graph, and n.

D.3 Methods, steps, and oracle accounting

Table 3 lists six accelerated/nonaccelerated pairs and five baselines. DGG-Prox and DGG-STL are from Domke et al. (2023); the other baselines are Price-SPGD (Kim et al., 2026), SVRGVI (Luu et al., 2025), and FB-GVI (Diao et al., 2023). All optimize the same full-covariance Gaussian objective. Write $s_p = \sqrt{\max\{\kappa, 4\}}$, $s_e = \sqrt{\max\{5\kappa, 4\}}$, and $r_b = 1 + (d+2)/b$.

The proximal and Cholesky projected variants use $\rho = (1 - s^{-1})/(1 + s^{-1})$, with $s = s_p$ and $s = s_e$, respectively; STL and Price-STL use $s = \sqrt{15\kappa}$. Nonaccelerated variants use the same estimator, step, and feasible set with no momentum. There are no restarts or step-size searches. Price-STL uses a score-corrected mean estimator, whereas Price-SPGD uses the raw mean gradient. SVRGVI uses

Method	Oracle	Coord.	Entropy treatment / domain	Step η
FD-prox (Alg. 1)	1st	L	Entropy and factor-ball prox	$1/(\alpha s_p^2)$
FD-proj (Alg. 4)	1st	L	Entropy extension, K_L	$1/(\alpha s_e^2)$
STL-proj (Alg. 2)	1st	C	Full gradient, K_C	$1/(15\beta)$
Paired-prox (Alg. 1)	2nd	L	Entropy and factor-ball prox	$1/(\alpha s_p^2)$
Paired-proj (Alg. 4)	2nd	L	Entropy extension, K_L	$1/(\alpha s_e^2)$
Price-STL-proj (Alg. 2)	2nd	C	Full gradient, K_C	$1/(15\beta)$
DGG-Prox	1st	L	Entropy prox, $L_{ii} > 0$	$1/(\beta r_b)$
DGG-STL	1st	C	Full gradient, $C \succeq \beta^{-1/2} I$	$1/(2\beta r_b)$
Price-SPGD	2nd	L	Entropy prox, $L_{ii} > 0$	$1/\beta$
SVRGVI ($c = 1$)	2nd	Σ	BW forward–backward step	$1/\beta$
FB-GVI	2nd	Σ	BW forward–backward step	$1/(2\beta)$

Table 3: **Methods and fixed steps.** Each of the first six rows includes accelerated and nonaccelerated variants. L is a positive-diagonal Cholesky factor and C is symmetric positive definite. Oracle order refers to target derivatives.

$c = 1$. Covariance-based methods average the b derivative estimates before one forward–backward update.

Per update, FD uses $d + 1 + b$ gradients, including $d + 1$ calibration queries. Paired uses b gradients and db HVPs at independent gradient and Hessian locations. Price-STL, Price-SPGD, SVRGVI, and FB-GVI have the same counts but share derivative locations. STL, DGG-Prox, and DGG-STL use b gradients and no HVPs.

D.4 OpenML execution and objective evaluation

All methods start from $q_0 = \mathcal{N}(0, I_d)$ and run for 600 updates, with checkpoints $0, 10, \ldots, 600$ and five seeds $20260831 + 997j$, $j = 0, \ldots, 4$. We use $b = 1, 8, 64, 256$. Batches consist of Gaussian Monte Carlo samples, each using the full likelihood. For STL and Price-STL, these practical choices do not enforce the sufficient batch bounds in Theorem 29.

Objective and numerical optimum. Each likelihood term depends on a one-dimensional Gaussian marginal. We evaluate the prior and entropy analytically and the likelihood by Gauss–Hermite quadrature. Quadrature is used only for evaluation and numerical references. The plotted gaps subtract a common reference $F_{\mathrm{ref}} \approx F^\star$, obtained by solving the stationary equations in (m, h) with $\Sigma = (Q_\tau + \operatorname{diag} h)^{-1}$.

References require a Cholesky-coordinate gradient norm at most 10^{-5}, a covariance stationary-equation residual at most 10^{-8} in relative norm, and agreement between quadrature orders 192 and 384 within 10^{-8} for the objective and 10^{-5} for the gradient. Checkpoints use successive order pairs $(192, 384)$, $(384, 768)$, and $(768, 1536)$ until objective agreement is within 10^{-8}; each gap subtracts the reference at the same order. These are numerical consistency checks, not rigorous integration-error bounds.

Curves and hitting times. Graph-probit figures show five-seed arithmetic means with empirical 20–80 percentile bands; solid and dashed lines distinguish accelerated and nonaccelerated variants. All five evaluations are required at each plotted checkpoint. For mean gap $\overline{g}_k$ and common initial gap g_0, the hitting time at relative threshold $\delta \in (0, 1)$ is the first recorded checkpoint with $\overline{g}_k \le \delta g_0$. It is measured on the ten-update grid and right-censored if the threshold is not reached by update 600; a finite crossing wins over a missing crossing.

D.5 Graph-probit comparisons and batch sensitivity

At $b = 64$, accelerated FD-proj, Paired-proj, and Price-STL-proj reach the 1% threshold earlier on every dataset (Table 4). The proximal pairs tie at the first checkpoint, $k = 10$, at this threshold.

Table 4 reports all six pairs at four thresholds.

Matched pair	10^{-2}	10^{-3}	10^{-4}	10^{-5}
FD-prox	0/0/5/0	3/0/2/0	4/0/1/0	5/0/0/0
FD-proj	5/0/0/0	4/0/1/0	5/0/0/0	5/0/0/0
STL-proj	2/3/0/0	2/2/0/1	2/0/0/3	1/0/0/4
Paired-prox	0/0/5/0	3/0/2/0	4/0/1/0	5/0/0/0
Paired-proj	5/0/0/0	4/0/1/0	5/0/0/0	5/0/0/0
Price-STL-proj	5/0/0/0	5/0/0/0	5/0/0/0	5/0/0/0

Table 4: **Graph-probit hitting-time comparisons at $b = 64$.** Thresholds are fractions of the initial mean gap. Entries count datasets with an earlier accelerated crossing / earlier nonaccelerated crossing / equal recorded crossing / no crossing by either variant. Matched-pair rankings are identical under derivative-query budgets.

Query costs. Figure 6 replots the $b = 64$ curves against cumulative gradient queries. FD includes its $d + 1$ calibration gradients per update, shifting its curves relative to STL and the gradient-only baselines. In the second-order comparisons, all methods use b gradients and db HVPs per update. Thus the within-pair momentum comparisons hold under both iteration and derivative-query budgets, while costs across estimators differ.

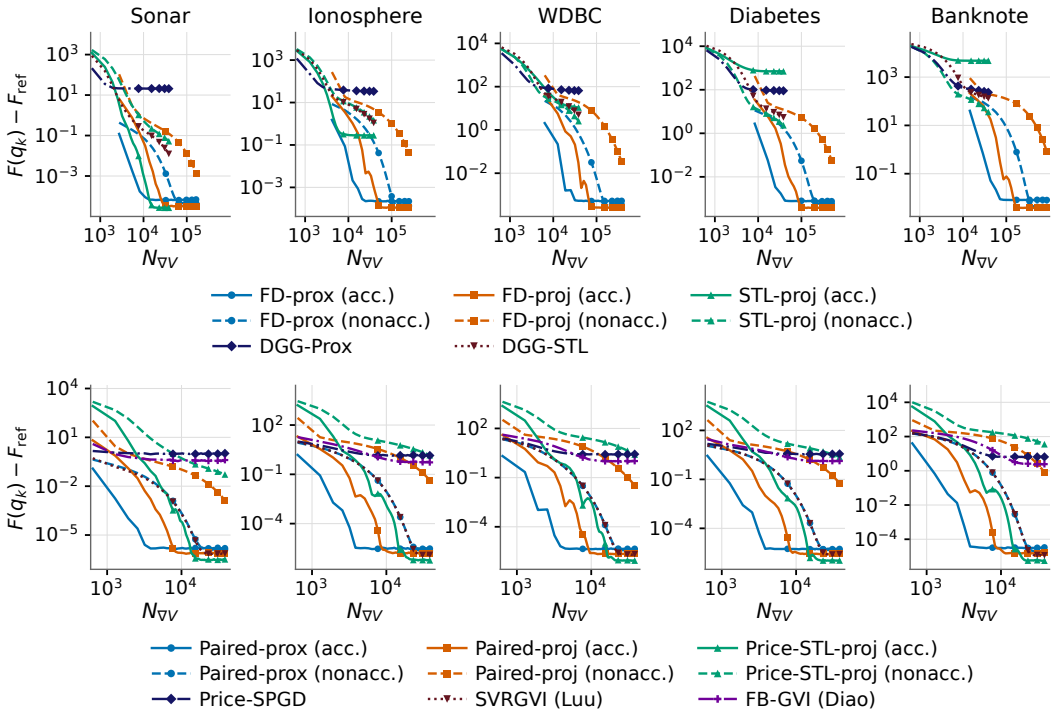


Figure 6: **Graph-probit convergence by gradient-query cost at $b = 64$.** Rows show first- and second-order methods on all five datasets. Means and bands follow Figure 3; the logarithmic horizontal axis uses cumulative gradient counts at $k = 10, 20, \ldots, 600$, including calibration and setup queries. First-order methods use no HVPs. For second-order methods, $N_{\mathrm{HVP}} = d[N_{\nabla V} - N_{\nabla V}(0)]$, where the setup count $N_{\nabla V}(0)$ is 1 for Paired-prox and 0 otherwise.

Increasing the batch size substantially improves accelerated STL-proj (Table 5). At $b = 256$, it reaches the 1% threshold in 30 updates on the first four datasets and 40 on Banknote, versus 50 and 60 for the nonaccelerated method, and has a smaller final gap on every dataset. Figure 7 shows all seventeen methods at this batch size. Accelerated Price-STL-proj already achieves final mean gaps below 6×10^{-6} on all datasets at $b = 64$.

	$b = 8$		$b = 64$		$b = 256$	
Dataset	Acc.	Nonacc.	Acc.	Nonacc.	Acc.	Nonacc.
Sonar	1.01×10^{3}	5.66×10^{-2}	2.70×10^{-5}	5.24×10^{-2}	2.84×10^{-6}	5.24×10^{-2}
Ionosphere	3.00×10^{3}	9.27×10^{1}	2.84×10^{-1}	1.46×10^{0}	1.09×10^{-5}	1.46×10^{0}
WDBC	9.61×10^{3}	1.28×10^{3}	1.09×10^{1}	2.56×10^{0}	3.34×10^{-5}	2.55×10^{0}
Diabetes	1.90×10^{4}	9.95×10^{3}	7.03×10^{2}	2.33×10^{0}	7.35×10^{-5}	2.32×10^{0}
Banknote	4.93×10^{4}	3.71×10^{4}	4.78×10^{3}	3.61×10^{1}	1.09×10^{0}	3.57×10^{1}

Table 5: **STL batch sensitivity.** Five-seed mean gaps $F(q_{600}) - F_{\mathrm{ref}}$, with identical stepsizes and initialization across batches.

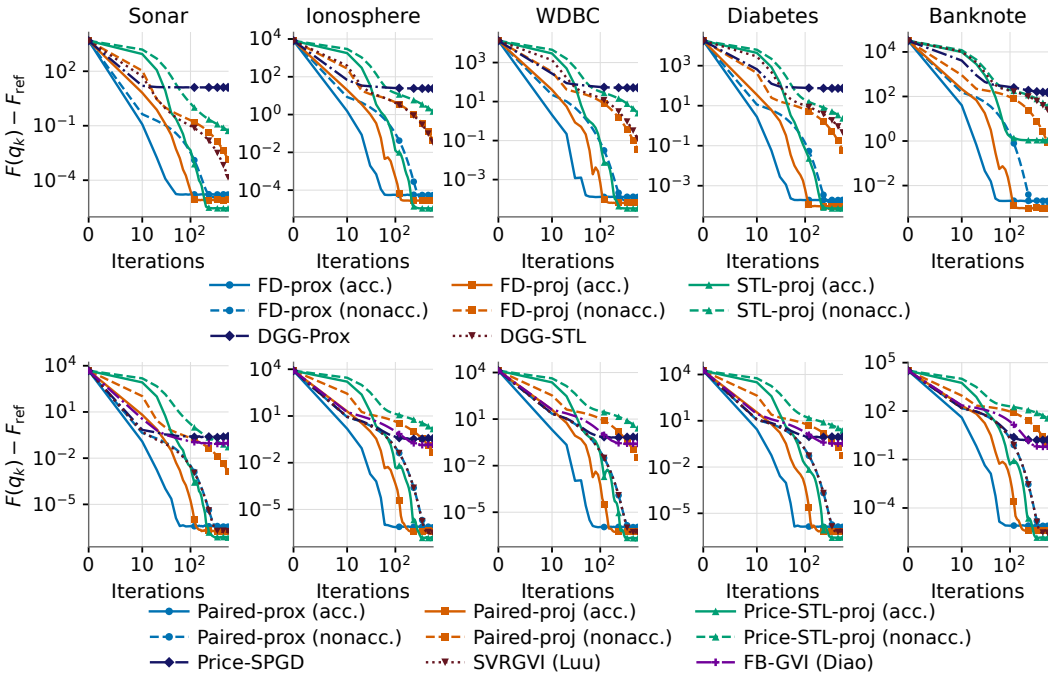


Figure 7: **Graph-probit comparisons at $b = 256$.** Rows show first- and second-order methods on all five datasets. Plotting conventions follow Appendix D.4; other settings match Figure 3.

D.6 LOGISTIC-REGRESSION COMPARISONS

To complement the graph-probit results, we also report experiments with Bayesian logistic-regression posteriors.

Setup and evaluation. We use the same five datasets and preprocessing as Appendix D.2. With $X = [X_0, \mathbf{1}_n]$ and $d = p + 1$, a standard Gaussian prior on all coefficients, including the intercept, gives

$$V_{\mathrm{logit}}(w) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n}\big[\log(1 + e^{x_i^\top w}) - y_i x_i^\top w\big], \qquad \alpha = 1, \quad \beta = \kappa = 1 + \tfrac{1}{4}\|X\|_{\mathrm{op}}^2.$$

The dimensions are $61, 34, 31, 9, 5$ and the condition-number bounds are $635.81, 774.27, 1890.31, 403.12, 748.71$ for Sonar, Ionosphere, WDBC, Diabetes, and Banknote, respectively. We compare all seventeen methods at $b = 1, 8, 64$, using the steps in Appendix D.3 and the initialization, seeds, update budget, and plotting conventions in Appendix D.4. Logistic expectations use Gauss–Hermite quadrature when the marginal standard deviation is below 0.75. Otherwise, we integrate the positive part analytically and the remaining softplus correction by Gauss–Legendre quadrature on $[0, 40]$. We compute F_{ref} by L-BFGS-B over the mean, log-diagonal, and strict lower-triangular entries of L, requiring a Cholesky-coordinate gradient norm at most 10^{-5} and the objective/gradient quadrature-consistency tolerances in Appendix D.4.

Results. Across five datasets and $b = 1, 8, 64$, acceleration reaches the 1% threshold earlier in 65 of 90 matched comparisons (Table 6). Figures 8 and 9 show every method, including the five baselines.

Matched pair	Acc. earlier	Nonacc. earlier	Same checkpoint	Neither reached
FD-prox	7	7	1	0
FD-proj	10	1	2	2
STL-proj	12	1	0	2
Paired-prox	9	3	3	0
Paired-proj	12	0	3	0
Price-STL-proj	15	0	0	0

Table 6: **Logistic-regression hitting-time comparisons.** Each row covers five datasets and three batches, using a hitting threshold of 1% of the initial mean gap.

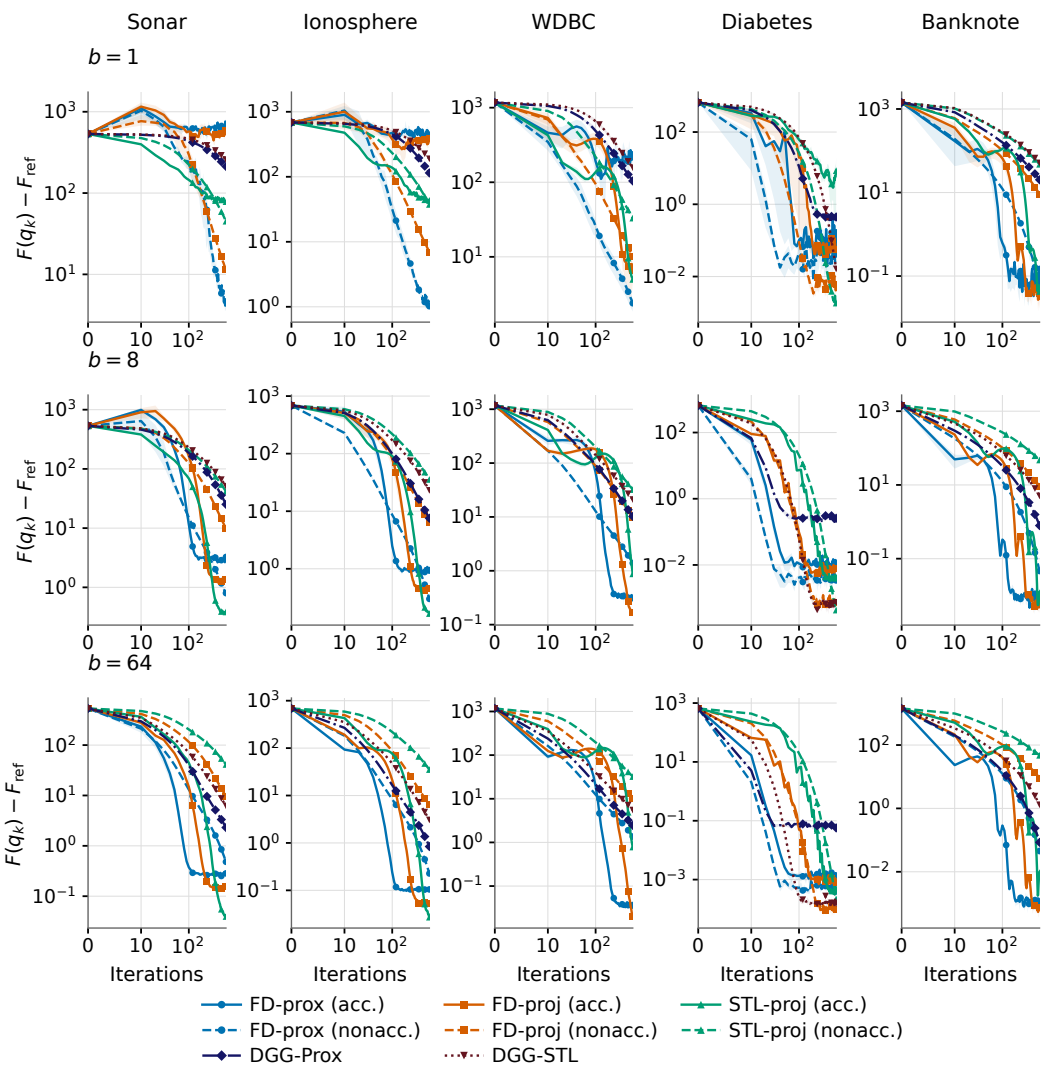


Figure 8: **First-order methods on logistic-regression posteriors.** Rows show $b = 1, 8, 64$; columns show the five datasets. Plotting conventions follow Appendix D.4.

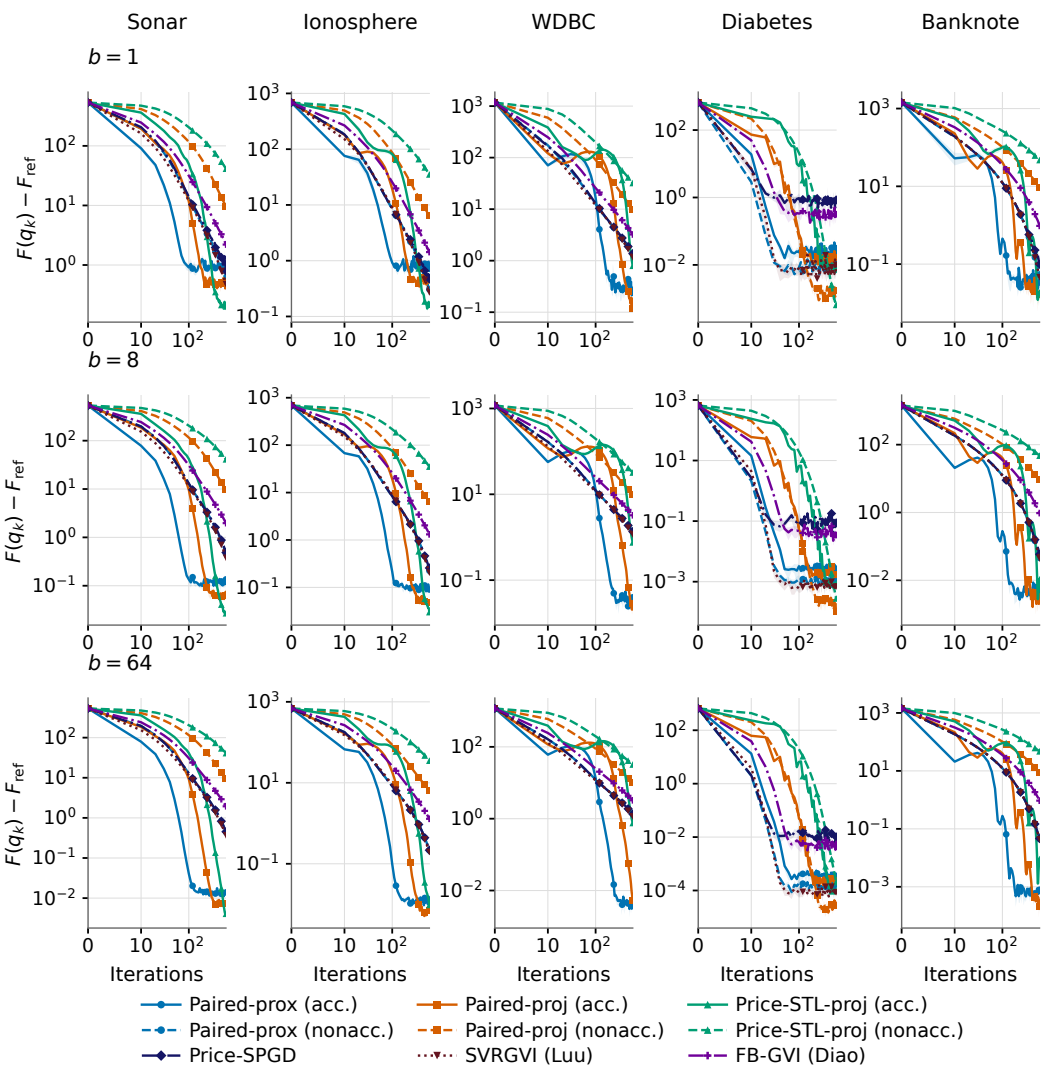


Figure 9: **Second-order methods on logistic-regression posteriors.** Rows show $b = 1, 8, 64$; columns show the five datasets. Plotting conventions follow Appendix D.4.