Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Policy-Risk Analysis and Feedback Allocation for Budgeted Frozen-Expert Routing

Yinpu Ma

mayinpu@163.com

North Minzu University

Research Article

Keywords: expert routing, finite-population sampling, policy risk, Riemannian optimization, feedback allocation, polar retraction

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11636351/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Policy-Risk Analysis and Feedback Allocation for Budgeted Frozen-Expert Routing

Anonymous author(s)

Abstract

A frozen expert library can support contextual routing without modifying its members, but acquiring the feedback needed to train the router may be costly. A finite response table introduces a specific statistical issue: previously observed outcomes can be reused, whereas rereading one entry does not produce an independent response. We study this setting for a categorical router with contrast Stiefel coordinates. We construct a cache-corrected estimator that remains conditionally unbiased after pilot-dependent allocation and derive its finite-population covariance. The policy differential converts this covariance into an observable allocation criterion. An explicit finite-step bound then separates linearized probability variance from the nonlinear effects of the polar update. Feature-computable upper bounds and integer allocation controls delimit what can be certified with unknown moments. The analysis distinguishes these results from classical square-root allocation and from representation compatibility. Controlled finite-table calculations and experiments with five frozen classifiers on three public datasets evaluate risk, pilot overhead, feedback use and held-out utility. Uniform and common-context sampling, gradient-space allocation and payoff regression provide controls. At the reporting budget, pilot-based policy allocation has 1.77–1.98 times the leading risk of no-pilot uniform sampling in four ordinary fixtures; a constructed concentration example reverses the comparison. These findings establish no universal utility advantage and do not constitute language-model service experiments.

Keywords: expert routing, finite-population sampling, policy risk, Riemannian optimization, feedback allocation, polar retraction

1 Introduction

An expert router selects one member of a fixed computational library for each input. Its purpose is to use variation in expert competence without paying for every expert at deployment. Examples include selecting a predictive model, choosing a reasoning

intervention, and allocating requests to heterogeneous services. These applications differ in their accessible features, feedback and resource costs (Yin and Zhang 2026; Chadli et al. 2026). In each case, an inexpensive deployment rule may require a substantial amount of offline information. The cost of acquiring that information should therefore be part of the learning problem.

We consider feedback stored in a finite table. Each row is a context and each column is an expert; an entry records the outcome obtained by applying that expert to that context. Once observed, the outcome can be reused at subsequent router states. The source of sampling randomness is the choice of entries to observe, rather than an assumed new response every time a stored value is read. This distinction is particularly relevant when studying released model outputs or deterministic frozen predictors. It changes the estimator, the accounting convention and the interpretation of repeated experiments.

Two questions arise. First, how can a sampled update use an expanding cache while retaining a precise relation to the full-information gradient? Pilot observations can influence subsequent sample counts, so simply averaging all observed values is not automatically conditionally unbiased. Second, which estimation error should an allocator minimize? A gradient-space norm treats all tangent directions alike. Their effects on routing probabilities need not be alike: the input features and the softmax differential can attenuate some directions and retain others. Thus gradient variance is a useful control, but it need not order observable policy error.

Classical stratified sampling already supplies variance decompositions and square-root allocations (Neyman 1934; Carpentier et al. 2015). Adaptive bandit sampling likewise has established variance-reduction methods (Zhao et al. 2025). Our aim is not to introduce a new allocation principle under a different name. We instead examine what remains valid after conditioning on cached feedback and pilots, how the covariance enters an actual finite router update, and how much information an unknown-moment allocator must acquire to use that relation. A lower probability error is also distinct from higher task utility; the latter requires a separate evaluation.

The router uses fixed-temperature linear-softmax probabilities in contrast Stiefel coordinates. This choice makes the policy map and feasible update explicit, allowing all derivatives and constants to be specified. Its geometry is standard (Edelman et al. 1998; Absil et al. 2008). Matched geometric representations can support equivalent algorithms (Zheng et al. 2025), while landscape correspondence under a smooth lift does not fix an arbitrary discrete update (Levin et al. 2025). We therefore use representation compatibility as an implementation control. The principal risk compares sampled and full-information updates in the *same* policy space, so that it remains meaningful after a compatible implementation removes representation mismatch.

The paper makes three contributions.

1. We formulate cache-aware finite-table feedback and construct an estimator that is conditionally unbiased after pilot-dependent production counts. Its covariance includes the observed contribution exactly and retains the remaining-population correction. This specifies the conditions under which cached outcomes can be reused without modeling them as independent generations.

2. We derive a finite-step policy-error bound from the covariance of the canonical update and the policy differential. We give a feature-computable upper bound for unknown moments and controls for integer allocation. The classical allocation algebra, the plug-in implementation and the certified quantities are stated separately.
3. We evaluate the resulting distinctions through finite-table calculations and actual frozen-expert predictions on three public datasets. The comparison retains pilot charges, common-context sampling, development-only selection and a nongeometric payoff-regression baseline. It reports unfavorable outcomes as well as successful mechanism checks, making the cost of adaptivity visible.

The theoretical results concern a specified update at a fixed, history-measurable state. They do not imply a global convergence or contextual-bandit regret guarantee. The experiments address finite-table routing with frozen CPU predictors; they establish neither language-model reasoning gains nor a general preference for orthogonality. This scope permits a direct connection between a computing task, the feedback available to its learner, and the numerical quantities used to assess its update.

2 Related work and contribution boundaries

2.1 Routing with a fixed action library

Routing can select a model, an intervention, or an expert; these choices expose different feedback and computational costs. Yin and Zhang (2026) select among answering, repeated sampling, retrieval and deferral for a fixed language model. Their behavioral features require candidate generations, and their full-signature protocol returns one of the results already generated during feature extraction. This is a different information interface from a query-feature router trained through restricted feedback. It also shows that selecting an intervention for a frozen model is an established application, rather than a contribution of the present work. Energy-aware model selection further requires explicit quality and resource models (Chadli et al. 2026); response count, historical monetary cost and measured energy are distinct quantities.

Our action library and input features remain fixed. The learning problem concerns which feedback entries to acquire and how their estimation error affects the updated policy. A stored expert outcome is a fixed table entry: reading it again does not create an independent response. This distinction permits evaluation with frozen experts, while preventing a finite response table from being interpreted as a repeated-generation oracle. The experiments compare methods under their stated information budgets and separate acquisition, optimization and deployment costs.

2.2 Contextual allocation and partial feedback

Li et al. (2026) allocate simulations across designs and contexts using a quadratic response model and Bayesian myopic criteria. They distinguish fixed-budget selection quality from fixed-precision stopping and prove asymptotic consistency under their modeling assumptions. Their posterior stopping criterion is not a finite-budget frequentist guarantee. We adopt neither their quadratic response model nor independent

repeated simulation at each table entry; the connection is the allocation of limited offline information to a contextual decision rule.

Stage-wise constrained bandits (Pacchiano et al. 2025) impose a per-round cost condition in probability or expectation while controlling cumulative regret. Our reservation of a finite feedback budget has a different purpose and does not imply such a regret guarantee. Likewise, Ban et al. (2026) learn signed reward residuals with an exploration network and analyze online contextual regret. Their result depends on a neural function class and overparameterized training assumptions. It does not establish convergence or regret for an orthogonal softmax router. A baseline adapted to our feedback interface is therefore distinguished from a reproduction of the original online algorithm.

2.3 Sampling variance and observable policy error

Adaptive variance reduction is established. Zhao et al. (2025) use online stochastic mirror descent to select clients from bandit feedback, explicitly noting that the sampling principle extends beyond federated learning. Inverse-inclusion estimation is classical (Horvitz and Thompson 1952), as are stratified square-root allocation rules (Neyman 1934; Carpentier et al. 2015). Replacing a gradient-norm weight by a policy-sensitivity weight does not by itself supply a new allocation principle.

The issue studied here is the finite-step probability error after the sampled update. Its leading term can be expressed through the covariance of the acquired gradient contributions and the policy differential. The nonlinear update, incomplete observations and integer budget must then be treated separately. We compare this observable criterion with gradient-space variance and uniform allocation, retaining pilot costs and the conditions needed for unbiased estimation. A reduction in the proposed risk is not identified with higher utility or an improved global convergence rate.

2.4 Representation, geometry and the chosen update

The canonical Stiefel metric and matrix retractions are standard (Edelman et al. 1998; Absil et al. 2008; Absil and Malick 2012; Boumal 2023). Smooth parametrizations can preserve landscape properties (Levin et al. 2025), while matched metrics can make algorithms in different representations equivalent. For example, Zheng et al. (2025) establish conjugate-gradient equivalences for fixed-rank Hermitian positive-semidefinite models and relate performance to metric-dependent conditioning. Consequently, neither orthogonal coordinates nor a general assertion that representation affects optimization is claimed as new.

Our contrast coordinates remove a policy-invisible completion. Compatible updates provide a representation control; independently applied polar updates can still differ at finite step size. That discrepancy motivates an implementation audit, whereas the principal risk compares a noisy and a full-information update in the same policy space. This keeps the sampling question meaningful even when the representation discrepancy has been removed. General stochastic retraction guarantees (Davis et al. 2025) and sampling analyses for stochastic trust-region methods (Ha et al. 2025) supply useful boundaries, but their objectives and oracle assumptions are not transferred to our router.

Manifold-valued approximation offers a complementary example of separating a tangent-space construction from its observable output error. Wang et al. (2026) combine several tangent-space approximations through weighted Fréchet means for parametric model reduction. Their method assumes given samples and does not choose a feedback allocation. We use this distinction to organize the analysis and numerical diagnostics, rather than importing their approximation guarantees.

3 Budgeted learning from fixed experts

3.1 Feedback and the local model

The router-training contexts have fixed features $h_i \in \mathbb{R}^d$, $i = 1, \dots, N$, with $\|h_i\|_2 \le H$. A fixed library of m experts defines rewards $r_{ij} \in [0, 1]$. Revealing an initially hidden entry costs one unit of training feedback. Its value can then be cached and reused. We count newly revealed entries, rather than iterations or repeated reads. A cached correctness score is not a new independent generation from a language model.

Let $c_j \in [0, 1]$ be a known deployment-cost scenario and fix $\lambda \ge 0$. For $r = m - 1$, choose a contrast matrix $C \in \mathbb{R}^{m \times r}$ with $C^\top C = I_r$, $C^\top \mathbf{1} = 0$, and $CC^\top = I_m - \mathbf{1}\mathbf{1}^\top/m$. Define

$$p_U(h) = \mathrm{softmax}(CU^\top h/\tau), \quad U \in \mathrm{St}(d, r), \quad \tau > 0, \tag{1}$$

$$F(U) = \frac{1}{N} \sum_{i,j} p_U(j \mid h_i)(r_{ij} - \lambda c_j). \tag{2}$$

Temperature is fixed within each fit. Orthogonality restricts the logit scale and capacity; no accuracy advantage is assumed. The penalty is a utility trade-off, not a hard deployment-cost constraint. Feedback acquisition, prescribed deployment cost and measured local computation time are reported separately.

For the canonical Stiefel metric, write

$$\langle \xi, \zeta \rangle_{c,U} = \mathrm{tr}\{\xi^\top (I - \tfrac{1}{2} UU^\top)\zeta\}, \qquad \mathcal{L}_U(A) = A - UA^\top U. \tag{3}$$

With e_j a standard basis vector, the local derivatives are

$$A_{ij} = \frac{p_U(j \mid h_i)}{\tau} h_i \{C^\top (e_j - p_U(h_i))\}^\top, \qquad D_{ij} = r_{ij} \mathcal{L}_U(A_{ij}), \tag{4}$$

$$g_{\mathrm{cost}} = -\frac{\lambda}{N} \sum_{i,j} c_j \mathcal{L}_U(A_{ij}), \qquad g = \mathrm{grad}_c\, F(U) = g_{\mathrm{cost}} + \frac{1}{N} \sum_{i,j} D_{ij}. \tag{5}$$

The cost gradient is known without revealing rewards. These formulas differentiate the router, not an expert's parameters. The update is the established polar retraction

$$T_U(\eta, \xi) = \mathrm{Retr}_U(\eta\xi) = (U + \eta\xi)(I + \eta^2 \xi^\top \xi)^{-1/2}, \quad \xi \in T_U\, \mathrm{St}(d, r). \tag{6}$$

3.2 Conditional estimation with a persistent cache

Condition on the current history $\mathcal{H}$, fixing U and the observed index sets K_j. Let $O_j = \{1, \dots, N\} \setminus K_j$ and $M_j = |O_j|$. Draw a uniform pilot $P_j \subseteq O_j$ of size k_j without replacement; its size is fixed before reading outcomes. After reading the pilots, choose production counts n_j. Write $Q_j = O_j \setminus P_j$ and $R_j = |Q_j|$. For every $R_j > 0$, draw a uniform subset $S_j \subseteq Q_j$ of size $1 \le n_j \le R_j$, independently across experts conditional on the pilots. Counts can use pilot values but not production values. Use

$$\widehat{g} = g_{\text{cost}} + \frac{1}{N} \sum_j \left\{ \sum_{i \in K_j \cup P_j} D_{ij} + \frac{R_j}{n_j} \sum_{i \in S_j} D_{ij} \right\}. \tag{7}$$

The last sum is zero for $R_j = 0$. Including all known and pilot contributions is essential: the remaining-population mean alone does not estimate the original population mean without bias. Production rewards enter the cache after forming this estimator. The observed-total plus inverse-inclusion construction belongs to classical finite-population estimation (Horvitz and Thompson 1952). Its role here is to specify the conditional estimator used after history-dependent caching and pilot-dependent count selection.

Reserve one production observation in each nonempty unknown stratum before purchasing pilots. The implementation purchases up to a fixed pilot cap per expert in round-robin order, retaining that reservation. Each block costs $\sum_j (k_j + n_j)$ new entries. If the budget cannot fund a complete all-strata block, no incomplete estimator is used. When the full table is cached, exact gradients require no new feedback, but their local computation remains part of the runtime record.

3.3 Allocation using observable probability changes

Fix design features $\mathcal{V} = \{\widetilde{h}_\ell\}_{\ell=1}^{L}$ from the training inputs before examining final-test outcomes. Define

$$J_{U,h}[\xi] = \tau^{-1}[\operatorname{diag}(p_U(h)) - p_U(h)p_U(h)^\top] C \xi^\top h, \tag{8}$$

$$\mathcal{T}_U(\xi) = L^{-1/2}(J_{U,\widetilde{h}_1}[\xi]^\top, \dots, J_{U,\widetilde{h}_L}[\xi]^\top)^\top. \tag{9}$$

This measures changes visible through the features and softmax. A narrow design distribution can miss relevant directions; it does not guarantee generalization. For $R_j \ge 2$ let

$$w_j = \left(\frac{R_j}{N}\right)^2 \frac{1}{R_j - 1} \sum_{i \in Q_j} \|\mathcal{T}_U(D_{ij} - \overline{D}_j)\|_2^2, \quad \overline{D}_j = R_j^{-1} \sum_{i \in Q_j} D_{ij}; \tag{10}$$

set $w_j = 0$ for $R_j \le 1$. Theorem 4.1 identifies the linearized risk as $\eta^2 \sum_j w_j (1/n_j - 1/R_j)$. True weights require unseen rewards.

We implement a regularized *policy-covariance* plug-in rule and a *policy-safe* alternative. The former uses fresh-pilot sample covariance after applying $\mathcal{T}_U$, multiplied

Algorithm 1 One cached finite-population router update

Require: U, cache (K_j), quota b, pilot cap k, design features, $\eta > 0$

Reserve one production entry per nonempty O_j
if the reservation exceeds b **then**
    return no update and an incomplete-block flag
end if
Choose pilot counts before seeing outcomes; reveal uniform pilots
Compute plug-in weights or pilot-safe upper weights
$B_{\mathrm{prod}} \leftarrow \min\{b - \sum_j k_j, \sum_j R_j\}$
Solve (11) by marginal allocation
Draw uniform production sets independently across experts
Reveal production rewards and form (7)
$U_+ \leftarrow T_U(\eta, \widehat{g})$; update cache and resource record
return U_+, cache and recorded new entries, pilots and local time

by $(R_j/N)^2$. Its ridge is 0.01 times the feature-only squared-range bound. With fewer than two pilots, a feature-only bound replaces the sample covariance. This is a heuristic estimate, not a confidence bound. The safe alternative uses Proposition 4.4. Gradient-covariance allocation replaces $\mathcal{T}_U$ by the identity. Uniform controls either buy no pilots or buy the same pilots.

For supplied nonnegative weights $\widetilde{w}_j$ solve

$$\min_{n} \sum_{j:R_j>0} \widetilde{w}_j(1/n_j - 1/R_j), \quad 1 \le n_j \le R_j, \quad \sum_j n_j = B_{\mathrm{prod}}. \tag{11}$$

Starting with one per active stratum, assign each additional unit to the eligible expert with greatest $\widetilde{w}_j/\{n_j(n_j+1)\}$. Ties have a fixed order; zero gains are resolved by balancing remaining fractions. This separable integer optimization and its continuous square-root analogue are standard allocation principles, not new optimization theorems.

Steps are selected on development data within the declared candidate budget. The certificates diagnose a specified positive update; they are not optimized by taking $\eta = 0$. A joint progress-constrained step controller is not implemented. The analysis adds cache conditioning, an observable policy-space target and a finite-step interpretation, rather than a universal preference for orthogonality or a claim that pilot overhead always pays.

4 Conditional covariance and finite-step policy risk

The reward table is fixed. Randomness comes from the sampled table indices, not unseen generations from an expert. Condition first on the history $\mathcal{H}$ and then on all fresh pilots and their outcomes, denoted $\mathcal{P}$. The current frame, design features, remaining populations and production counts are then fixed. All sums containing $1/n_j$ or $1/R_j$ below are restricted to $R_j > 0$; an exhausted stratum is evaluated exactly and contributes zero sampling covariance.

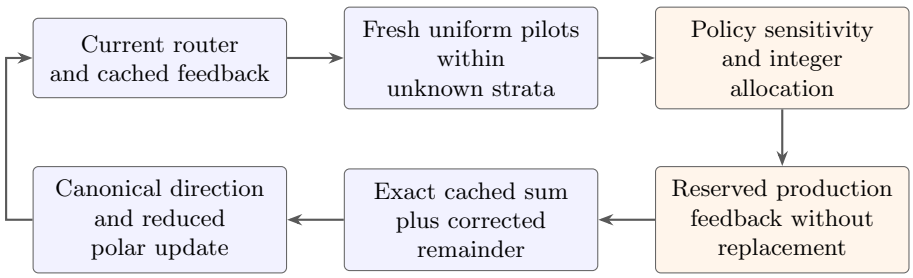

Fig. 1 Information flow for one cache-aware allocation block. New feedback is charged before the state update; known deployment costs use local probabilities and need no expert query. Full-information risk diagnostics are outside this learning loop.

4.1 Exact finite-population identities

Theorem 4.1 (Cache-corrected gradient and policy covariance)**.** *Under the conditional sampling law in Section 3, the estimator in (7) is tangent and*

$$\mathbb{E}[\widehat{g} \mid \mathcal{H}, \mathcal{P}] = g. \tag{12}$$

For $R_j \geq 2$ define

$$\Sigma_j = \frac{1}{R_j - 1} \sum_{i \in Q_j} \operatorname{vec}(D_{ij} - \overline{D}_j) \operatorname{vec}(D_{ij} - \overline{D}_j)^\top.$$

With zero contributions from $R_j \leq 1$,

$$\operatorname{Cov}(\operatorname{vec}(\widehat{g}) \mid \mathcal{H}, \mathcal{P}) = \sum_j \left(\frac{R_j}{N}\right)^2 \left(\frac{1}{n_j} - \frac{1}{R_j}\right) \Sigma_j, \tag{13}$$

$$V_g := \mathbb{E}[\|\widehat{g} - g\|_F^2 \mid \mathcal{H}, \mathcal{P}] = \sum_j \left(\frac{R_j}{N}\right)^2 \left(\frac{1}{n_j} - \frac{1}{R_j}\right) \operatorname{tr}(\Sigma_j), \tag{14}$$

$$V_p := \mathbb{E}[\|\mathcal{T}_U(\widehat{g} - g)\|_2^2 \mid \mathcal{H}, \mathcal{P}] = \sum_j w_j \left(\frac{1}{n_j} - \frac{1}{R_j}\right). \tag{15}$$

Moreover, $V_p \leq H^2 V_g / (4\tau^2)$.

Proof All contributions are tangent and the tangent space at fixed U is linear. The production sample mean has conditional expectation $\overline{D}_j$. Adding the known, pilot and expanded production sums therefore recovers the full gradient.

For a uniform subset of size n from R entries, let I_i be the inclusion indicators. Then $\mathbb{E}I_i = n/R$ and $\mathbb{E}I_iI_\ell = n(n-1)/\{R(R-1)\}$ for $i \neq \ell$. Expanding the covariance of the sample mean and using $\sum_i (D_i - \overline{D}) = 0$ gives $(1/n - 1/R)\Sigma$. Conditional independence across experts proves (13); taking its trace proves (14). Apply the same identity after the fixed linear map $\mathcal{T}_U$ to obtain (15). Finally the softmax Jacobian $S(p) = \mathrm{diag}(p) - pp^\top$ has norm at most $1/2$, so $\|J_{U,h}[\xi]\|_2 \leq \|h\|_2\|\xi\|_F/(2\tau)$. □

These identities allow U and the cache to depend on past rewards: that dependence is fixed by conditioning on $\mathcal{H}$. They do not permit using one fixed reward as several independent observations. The finite-population correction and the exact known sum are substantive differences from a repeated-generation model.

Proposition 4.2 (Integer optimum and weight sensitivity)**.** *Let $a = |\{j : R_j > 0\}|$ and $a \leq B_{\mathrm{prod}} \leq \sum_j R_j$. The marginal algorithm in Section 3 attains a global integer optimum for every supplied nonnegative weight vector. Let $n^\star$ minimize $V_w(n) = \sum_j w_j(1/n_j - 1/R_j)$ and let $\widehat{n}$ minimize $V_{\widehat{w}}$ on the same feasible set. Then*

$$0 \leq V_w(\widehat{n}) - V_w(n^\star) \leq \sum_{j:R_j>0} |w_j - \widehat{w}_j|(1 - 1/R_j). \tag{16}$$

If the common feasible set imposes $n_j \geq \ell_j$, replace $1 - 1/R_j$ by $1/\ell_j - 1/R_j$.

Proof Increasing n_j to $n_j + 1$ reduces the objective by $w_j/\{n_j(n_j+1)\}$. These gains are nonincreasing within a stratum. The largest available increments therefore select a largest-total-gain set of $B_{\mathrm{prod}} - a$ increments, because selecting an increment permits all preceding, no smaller increments in its stratum. This proves integer optimality. Optimality for $\widehat{w}$ gives

$$V_w(\widehat{n}) - V_w(n^\star) \leq \sum_j (w_j - \widehat{w}_j)(1/\widehat{n}_j - 1/n_j^\star).$$

Bound each reciprocal difference by the stated range. □

The continuous rule $n_j \propto \sqrt{w_j}$, when count constraints are inactive, follows from Cauchy–Schwarz and is classical. Proposition 4.2 does not assert accurate weight estimation: it separates that statistical error from solving the integer problem for the supplied weights.

Proposition 4.3 (The effect of observing common contexts)**.** *Consider an initially empty cache and no pilots. Compare n uniform contexts observed independently per expert with n uniform contexts whose rewards are observed for all experts. Both reveal mn entries. Put $Z_{ij} = \mathcal{T}_U(D_{ij})$ and let $S_{jk} = (N-1)^{-1}\sum_i (Z_{ij} - \overline{Z}_j)(Z_{ik} - \overline{Z}_k)^\top$. The linearized variances satisfy*

$$\begin{aligned} V_{\mathrm{ind}} &= (1/n - 1/N)\sum_j \mathrm{tr}(S_{jj}), \\ V_{\mathrm{row}} &= V_{\mathrm{ind}} + (1/n - 1/N)\sum_{j\neq k} \mathrm{tr}(S_{jk}). \end{aligned} \tag{17}$$

Thus there is no universal variance order between the two observation schemes. If every expert has reward one on every context, $V_{\mathrm{row}} = 0$, whereas V_{ind} need not vanish.

Proof For independent sampling the cross-expert covariances are zero. For common-context sampling, apply the finite-population mean covariance identity to $\sum_j Z_{ij}$ and expand its covariance. When all rewards equal one, $\sum_j D_{ij} = \mathcal{L}_U(\nabla_U \sum_j p_U(j \mid h_i)) = 0$ for every context. The common-context reward-gradient estimate is then exactly zero; the known cost gradient is deterministic. □

This elementary covariance identity explains why an independent-strata allocator must also be tested against shared-context acquisition. It does not extend (13) to dependent production samples without changing that equation. With a nonempty cache, costs and available blocks must be handled separately.

4.2 A pilot-dependent upper envelope

Proposition 4.4 (Observable upper weights)**.** *Fix $0 < \delta < 1$. For every $R_j \geq 2$ define b_j below, and define $\overline{s}_{P,j}$ when $k_j \geq 1$:*

$$b_j = \max_{i \in O_j} \|\mathcal{T}_U(\mathcal{L}_U(A_{ij}))\|_2^2, \qquad \overline{s}_{P,j} = \frac{1}{k_j} \sum_{i \in P_j} \|\mathcal{T}_U(D_{ij})\|_2^2. \tag{18}$$

For $k_j \geq 1$ put

$$u_j^{\mathrm{all}} = \min\left\{ b_j, \overline{s}_{P,j} + b_j \sqrt{\frac{\log(m/\delta)}{2k_j}} \right\}, \tag{19}$$

$$u_j^{\mathrm{rem}} = \max\{0, (M_j u_j^{\mathrm{all}} - k_j \overline{s}_{P,j})/R_j\},$$

$$\overline{w}_j = \left(\frac{R_j}{N}\right)^2 \frac{R_j}{R_j - 1} u_j^{\mathrm{rem}}. \tag{20}$$

For $k_j = 0$ use $u_j^{\mathrm{rem}} = b_j$; for $R_j \leq 1$ set $\overline{w}_j = 0$. Conditional on $\mathcal{H}$, with probability at least $1-\delta$ over the pilots, $w_j \leq \overline{w}_j$ for all experts. On that event every pilot-dependent feasible allocation has

$$V_p \leq \overline{V}_p(n) := \sum_{j : R_j > 0} \overline{w}_j (1/n_j - 1/R_j). \tag{21}$$

Proof Let $s_{ij} = \|\mathcal{T}_U(D_{ij})\|_2^2$. Since $r_{ij} \in [0,1]$, $0 \leq s_{ij} \leq b_j$. The Hoeffding bound for a uniform sample without replacement gives

$$\Pr\left\{ \frac{1}{M_j} \sum_{i \in O_j} s_{ij} > \overline{s}_{P,j} + b_j \sqrt{\frac{\log(m/\delta)}{2k_j}} \,\middle|\, \mathcal{H} \right\} \leq \delta/m.$$

This follows from the convex-order comparison of sampling without and with replacement and the bounded-variable exponential-moment inequality (Hoeffding 1963). The true mean

is also bounded by b_j. Subtracting the observed pilot sum bounds the remaining raw mean by u_j^{rem}. Furthermore,

$$\frac{1}{R_j-1}\sum_{i\in Q_j}\|\mathcal{T}_U(D_{ij}-\overline{D}_j)\|_2^2=\frac{R_j}{R_j-1}\left\{\frac{1}{R_j}\sum_{i\in Q_j}s_{ij}-\|\mathcal{T}_U(\overline{D}_j)\|_2^2\right\}\le\frac{R_j}{R_j-1}u_j^{\mathrm{rem}}.$$

A union bound proves the simultaneous envelope. All coefficients in (15) are nonnegative, which proves (21). Cases without pilots or with at most one remaining entry are deterministic. □

The confidence statement concerns the pilot event. On that event, (21) bounds a conditional expected risk over production sampling, not every realized error. Dropping the nonnegative squared mean makes the envelope conservative; small pilots can also make its range term dominate the data. Replacing $\mathcal{T}_U$ by the identity gives a corresponding gradient-variance envelope. For any nonnegative plug-in weight, the same event supplies $|w_j-\widehat{w}_j|\le\max\{\widehat{w}_j,|\overline{w}_j-\widehat{w}_j|\}$. This can be used in (16), but does not make pilot sample covariance unbiased for the remaining covariance.

4.3 Actual policy risk after a polar step

Lemma 4.5 (Uniform first-order remainder)**.** *Let $\|h\|_2\le H$, $s=H/\tau$ and $K_s=s/4+s^2$. For every tangent ξ and real η,*

$$\|p_{T_U(\eta,\xi)}(h)-p_U(h)-\eta J_{U,h}[\xi]\|_2\le K_s\eta^2\|\xi\|_F^2. \tag{22}$$

Proof For $p=\mathrm{softmax}(z)$ and $J=\mathrm{diag}(p)-pp^\top$, differentiating Ja along b yields

$$D^2\,\mathrm{softmax}(z)[a,b]=\mathrm{diag}(a)Jb-(p^\top a)Jb-p(a^\top Jb).$$

Since $\|J\|_2\le 1/2$, each term has norm at most $\|a\|_2\|b\|_2/2$; the bound $2\|a\|_2\|b\|_2$ is thus valid. The ambient policy map has derivative norm at most $s/2$ and second-derivative norm at most $2s^2$.

For tangent ξ, $(U+\eta\xi)^\top(U+\eta\xi)=I+\eta^2\xi^\top\xi$. Spectral calculus and $\sqrt{1+t}-1\le t/2$ imply

$$\|T_U(\eta,\xi)-U-\eta\xi\|_F\le\tfrac{1}{2}\eta^2\|\xi\|_F\|\xi\|_2.$$

The squared polar displacement equals $2\,\mathrm{tr}\{I-(I+\eta^2\xi^\top\xi)^{-1/2}\}$, hence $\|T_U(\eta,\xi)-U\|_F\le|\eta|\|\xi\|_F$. Taylor expansion of the ambient policy along this displacement therefore has an error at most $(s/4+s^2)\eta^2\|\xi\|_F^2$. □

Theorem 4.6 (Finite-step policy risk and allocation excess)**.** *Suppose $\|\widehat{g}\|_F,\|g\|_F\le K$ almost surely conditional on $\mathcal{H},\mathcal{P}$. Define*

$$\mathcal{R}_\eta(n)=\mathbb{E}\left[\frac{1}{L}\sum_{\ell=1}^{L}\|p_{T_U(\eta,\widehat{g})}(\widetilde{h}_\ell)-p_{T_U(\eta,g)}(\widetilde{h}_\ell)\|_2^2\,\middle|\,\mathcal{H},\mathcal{P}\right]. \tag{23}$$

For $\eta\ge 0$,

$$|\mathcal{R}_\eta(n)-\eta^2V_p(n)|\le\mathcal{E}_\eta(V_p(n)),$$

$$\mathcal{E}_\eta(v) = 4K_s\eta^3K^2\sqrt{v} + 4K_s^2\eta^4K^4, \tag{24}$$

$$\mathcal{R}_\eta(n) \le \{\eta\sqrt{V_p(n)} + 2K_s\eta^2K^2\}^2. \tag{25}$$

On the event of Proposition 4.4, replacing V_p by $\overline{V}_p$ yields a computable upper bound. For the true-weight integer optimum $n^\star$ and plug-in optimum $\widehat{n}$ on the same post-pilot feasible set,

$$\begin{aligned}\mathcal{R}_\eta(\widehat{n}) - \mathcal{R}_\eta(n^\star) \le \eta^2 \sum_{j:R_j>0} |w_j - \widehat{w}_j|(1 - 1/R_j) \\ + \mathcal{E}_\eta(V_p(\widehat{n})) + \mathcal{E}_\eta(V_p(n^\star)).\end{aligned} \tag{26}$$

The reference $n^\star$ minimizes linearized risk, not necessarily nonlinear risk or final utility.

Proof Apply Lemma 4.5 to $\widehat{g}$ and g. The stacked probability difference is $\eta\mathcal{T}_U(\widehat{g} - g) + e$ with $\|e\|_2 \le 2K_s\eta^2K^2$. Expand its squared norm and bound the cross term using Cauchy–Schwarz and $\mathbb{E}\|\mathcal{T}_U(\widehat{g} - g)\|_2^2 = V_p$. This proves (24) and (25). The latter increases with V_p, justifying the envelope substitution. Upper bound the risk at $\widehat{n}$, lower bound it at $n^\star$, and use (16) for the excess claim. □

A sufficient global bound is

$$K = \frac{2\sqrt{2}H}{\tau}(m + \lambda). \tag{27}$$

Indeed $\|\mathcal{L}_U(A_{ij})\|_F \le 2\sqrt{2}Hp_U(j \mid h_i)/\tau$. For each action, the known-plus-expanded-sample estimate is a convex combination of contributions bounded by $2\sqrt{2}H/\tau$. Summing over experts and bounding the exact cost gradient by $2\sqrt{2}\lambda H/\tau$ proves the claim. The resulting certificate may be loose, so tightness and absolute effect size must be reported together. Constants are not uniform as $\tau \downarrow 0$.

A tighter history-dependent bound can be obtained without new rewards. Let

$$\begin{aligned} g_{\text{known}} &= g_{\text{cost}} + \frac{1}{N}\sum_j \sum_{i \in K_j \cup P_j} D_{ij}, \\ K_{\text{loc}} &= \|g_{\text{known}}\|_F + \frac{1}{N}\sum_{j:R_j>0} R_j \max_{i \in Q_j} \|\mathcal{L}_U(A_{ij})\|_F. \end{aligned} \tag{28}$$

Every unknown reward lies in $[0, 1]$, so the triangle inequality shows that K_{loc} bounds both the exact remaining sum and every expanded production sum added to g_{known}. Theorem 4.6 can therefore use the smaller of (27) and (28). This refinement changes a certificate, not the training update; the numerical audits state which constant they evaluate.

Remark 4.7 (Utility and training scope). The risk compares sampled and exact empirical-gradient updates from the same current state. It does not compare independently evolved training trajectories or guarantee improved test utility. For a fixed utility vector $y \in [-\lambda, 1]^m$, $|y^\top(p-q)| \le \|y\|_2\|p-q\|_2$ connects policy discrepancy to a conditional utility discrepancy, but the exact-gradient step itself need not be optimal. The theory therefore complements, rather than replaces, independent tests of routing quality.

4.4 A representation-consistent implementation control

Proposition 4.8 (Compatible full-frame realization)**.** *Let $d \ge m$, $U^\top v = 0$, $\|v\|_2 = 1$ and $e_0 = \mathbf{1}/\sqrt{m}$. Then $\Phi = UC^\top + ve_0^\top \in \mathrm{St}(d,m)$ has policy p_U. For $U_+ = T_U(\eta,\xi)$ put*

$$v_+ = \frac{(I - U_+U_+^\top)v}{\|(I - U_+U_+^\top)v\|_2}, \qquad \Phi_+ = U_+C^\top + v_+e_0^\top. \tag{29}$$

Where the denominator is nonzero, Φ_+ is feasible and has policy p_{U_+} exactly. The condition $|\eta|\|\xi\|_F < 1$ is sufficient. At zero its derivative in $[C, e_0]$ coordinates is $[\xi, -U(v^\top\xi)^\top]$.

Proof Multiplying logits by $CC^\top$ subtracts their common mean. Projection and normalization make v_+ unit and orthogonal to U_+, so $\Phi_+^\top\Phi_+ = I_m$ and $\Phi_+C = U_+$. The polar displacement bound gives $\|U_+ - U\|_F \le |\eta|\|\xi\|_F$. Thus $\|U_+^\top v\|_2 < 1$ under the sufficient condition. The derivative of the projected completion at zero is $-U\xi^\top v$, orthogonal to v, so normalization preserves it. □

This specializes classical frame geometry, rather than defining a new global retraction. With matching canonical directions and shared feedback, the compatible full implementation has the same policy as the reduced implementation whenever all completions are defined. Thus sampling risk persists even after representation mismatch is removed. Metric-dependent algorithm equivalences have precedents (Zheng et al. 2025); arbitrary full and reduced polar steps or embedded-metric directions are not claimed to coincide.

5 Numerical evaluation

The evaluation separates three questions: whether the finite-population identities hold, whether their linearized criterion predicts finite-step probability error, and whether an allocator improves a budgeted routing pipeline. The first two admit full-information diagnostics; the third uses a development split for selection and an independent test split for reporting. None of the experiments calls a language-model service.

5.1 Data, experts and accessible feedback

We use the scikit-learn distributions of Digits, Breast Cancer Wisconsin (Diagnostic), and Wine (Pedregosa et al. 2011; Alpaydin and Kaynak 1998; Wolberg et al. 1993; Aeberhard and Forina 1992). Digits contains the bundled 1797-sample subset, rather

Table 1 Disjoint data partitions in every run. The raw feature dimension is distinct from the common 32-dimensional router input.

Dataset	Records	Raw dim.	Expert fit	Router train	Dev.	Test
Digits	1797	64	718	539	180	360
Breast Cancer	569	30	227	171	57	114
Wine	178	13	71	53	18	36

than all 5620 records in the UCI collection. The other datasets contain 569 and 178 records. These small public datasets provide different feature and class structures; they do not represent three independent production deployments.

For each dataset, ten stratified random splits partition the original records into expert fitting, router training, development and testing. The nominal proportions are 40/30/10/20%, with integer rounding performed by the specified splitting routine. Table 1 gives the resulting sizes. All splits are made before fitting any expert. The seed list is $11, 23, 37, 53, 71, 89, 101, 127, 149, 173$.

The fixed library contains Gaussian naive Bayes, a depth-three decision tree, logistic regression, five-nearest-neighbor classification, and a 100-tree random forest. Expert hyperparameters are fixed in the preprocessing program. Logistic regression and nearest neighbors use standardization fitted only on expert-fitting records. No expert is selected or fitted on router, development or test outcomes. The feedback entry y_{ij} is the correctness of expert j on record i. Conditional on the table, this entry is deterministic.

Router features are constructed independently of expert predictions. A standardizer fitted on the expert-fitting partition is followed by a fixed Gaussian projection to 31 coordinates and a constant coordinate. Each 32-vector is divided by the larger of its norm and one. The projection seed depends only on the dataset and is fixed across split seeds. Thus router features use no test labels, behavioral probe responses or paid embedding model.

We distinguish an acquisition unit from deployment cost. One newly revealed training-table entry consumes one acquisition unit; a previously revealed entry can be reused within the fit. The known deployment-cost scenario is proportional to $(1, 1.5, 2, 3, 6)$ in library order and is normalized to maximum one. These are explicit relative costs, not measured currency, energy or latency. Separate warmed CPU prediction timings are retained as implementation diagnostics. The experiment evaluates $\lambda = 0$ and 0.1 under this scenario and does not claim that the chosen cost vector models a particular provider.

5.2 Budgets, updates and comparison rules

Acquisition ceilings are $\lfloor fNm \rfloor$ for $f \in \{0.1, 0.25, 0.5, 0.8\}$ on each router-training table. All Stiefel fits use $d = 32$, $m = 5$, $\tau = 0.25$, and the same initial contrast frame within a split. There are at most 48 updates. A deterministic rule distributes the ceiling across approximately 20-entry blocks, reserving complete feasible production strata

before purchasing pilots. The program records unused quota, every newly observed entry and the cache size.

The final step grid is $(0.0625, 0.25, 1, 4, 16, 64, 256)$. The initial five-rate grid was expanded after development selection frequently reached its upper boundary; the original records are retained. Each fit is trained independently; the rate with highest exact development-table utility is selected, breaking ties by grid order. The full development response table is therefore a declared evaluation resource. Per-fit charges, their sum across candidate fits, and the union of revealed entries across fits are recorded separately. These are offline feedback-accounting quantities, not new API requests. Only selected predictors are evaluated on test outcomes.

The comparison includes uniform sampling without pilots; uniform production with paid pilots; gradient-covariance allocation; policy-covariance allocation; and the uncentered-moment upper-bound allocation. The pilot cap is two entries per active expert in each block, reduced when needed to retain production feasibility. For the upper-bound rule, $\delta = 0.05/(7T)$ at each update, where T is the number of planned blocks in that fit; a union bound covers its seven-fit family within a dataset/split/budget/penalty group, not all reported groups jointly. The numerical design distribution is a fixed uniform sample of up to 32 unlabelled router-training features, using seed $s + 30000$ for split seed s, shared by all candidates. A common-context control reveals complete expert rows and preserves cross-expert covariance. Its final incomplete-row allowance is retained as unused quota. The plug-in covariance rules use a fixed positive ridge and are not labeled confidence-valid variance estimators. The upper-bound rule is distinguished from them both in the equations and in the numerical records.

A separate payoff-regression baseline receives the same acquisition ceiling, samples feedback without replacement, and fits one ridge predictor for each expert. Its final regularization grid is $(0.0001, 0.001, 0.01, 0.1, 1, 10, 100)$ and is selected on development utility. It deploys the expert with largest predicted cost-adjusted payoff. This is a stronger, different policy family and deployment rule, included to assess whether the geometric router is useful relative to a simple computing baseline. It is not governed by the Stiefel risk theorem. Development-selected fixed-expert routing and the test-only best-expert-per-record reference further delimit achievable performance; the latter is unavailable to a deployed router.

Monitoring quantities never select a step, checkpoint or expert. Full-table gradient and objective diagnostics are explicitly logged as oracle work. They do not become training feedback. The protocol records software versions, table hashes, split indices and all candidate configurations. The repeated splits share the same source datasets; their variation is descriptive and does not create ten independent datasets.

5.3 Finite-population covariance and finite-step risk

The risk audit holds the current frame, cache and pilot entries fixed and resamples only the production context indices. It uses three fixed synthetic tables and one fixed training-table fixture from the Digits expert study. The synthetic tables have $N = 80$, $d = 8$ and $m = 4$; their rewards are smooth bounded values, one fixed binary table, and smooth values with a concentrated design distribution, respectively. The Digits

Table 2 Fixed-state allocation diagnostics at a new-feedback ceiling of $8m$. Entries are ratios of exact leading policy variances. Paid uniform uses the same fixed pilots as the adaptive rules; no-pilot uniform spends its entire allowance on production. The oracle is an unattainable diagnostic using complete-table moments. Ratios below one favor the numerator only for this stated fixed fixture.

Fixture	Budget	Policy/no-pilot	Policy/paid	Oracle/paid
Smooth table	32	1.771	0.943	0.927
Fixed binary table	32	1.897	0.997	0.954
Concentrated design	32	1.794	0.956	0.908
Digits training table	40	1.984	1.000	0.975

fixture has $N = 539$, $d = 32$ and $m = 5$. The temperature is 0.5. Design contexts are the first 24 training features, except that the concentrated fixture repeats the first four features six times. These contexts are declared numerical design points, not independent test examples.

The cache initially contains $j + 1$ entries for action j, with actions indexed from zero. The new-feedback ceilings are $2m$, $8m$ and $16m$. Uniform sampling uses no pilots. The other rules request four pilots per action, subject to the all-strata production reservation. The comparison also retains uniform production with the same paid pilots and an unavailable oracle allocation using the exact remaining-population moments. For each of 72 fixed-pilot configurations, 4,000 independent production subsets are generated and reused across six steps from 10^{-4} to 0.3125. A repeated table access does not represent an independent expert response.

Exact covariance is available to the audit because the complete table is known to the diagnostic program. The largest discrepancy between empirical gradient mean-square error and its exact finite-population value is 2.46 Monte Carlo standard errors. No recorded Monte Carlo mean exceeds the finite-step upper bound. This is a numerical check, not a proof of the bound. The safe moment event holds in the 12 policy-safe configurations; these observations do not establish its nominal probability coverage. The reported files retain direct Monte Carlo estimates and their standard errors. They additionally contain an audit-only control-variate estimate: the nonlinear squared probability error minus its samplewise linear term is averaged, then the exact linear-risk mean is restored. Its standard error describes production-subset Monte Carlo only.

A separate 8-context, 3-action fixture enumerates all 1,000 equally weighted production-subset combinations, with three known entries and two production entries per action. The gradient mean residual is 3.85×10^{-17}. The relative difference between the exact finite-step risk and its leading quadratic term grows from 1.52×10^{-6} at $\eta = 0.0002$ to 1.25×10^{-3} at $\eta = 0.2048$. This calculation has no Monte Carlo uncertainty and isolates the local approximation from sampling error.

Table 2 separates weight selection from pilot overhead. With the same paid pilots, policy weighting can reduce the leading risk relative to uniform production. It does not outperform the no-pilot control in these fixtures. At the smallest allowance the production reservation leaves too little pilot information to estimate covariance, while

Table 3 Constructed boundary diagnostic. Mean ratio of exact leading policy variance to the no-pilot uniform control over five fixed pilot sets; lower is better. Each ratio includes paid pilots. The oracle uses unavailable remaining-population moments. The construction isolates variance concentration and is not evidence of a general routing-performance advantage.

Pilots/action	Budget	Gradient weights	Policy weights	Safe weights	Oracle
2	64	0.513	0.532	0.557	0.217
2	128	0.408	0.427	0.449	0.085
8	64	0.557	0.637	1.024	0.447
8	128	0.256	0.324	0.594	0.149

the safe raw-moment weights remain conservative. Consequently, the audit supports the covariance identity and the need to charge pilots; it supplies no universal advantage of adaptive allocation. All 288,000 production-subset repetitions are numerical diagnostic work, reported separately from any trained router's feedback allowance.

A constructed variance-concentration boundary case.

To investigate when information spent on pilots can be amortized, we add one deliberately constructed diagnostic after the preceding audit: 160 copies of a normalized eight-dimensional feature, four actions, alternating fixed zero–one rewards for action zero, and constant reward one for the other actions. No fresh response is generated. The table has sampling variance in only one action stratum. Five fixed pilot seeds $11, 23, 37, 53, 71$, requested pilot counts two and eight, and feedback ceilings 64 and 128 are all retained. The seeds are five pilot realizations of one table, not five datasets.

In this concentrated construction, estimated allocation can recover its pilot overhead. Gradient-space weighting can be more effective than policy weighting, so the example does not establish an advantage of the policy-specific weights over every variance baseline. The safe raw-moment rule remains conservative: with eight pilots and budget 64, it improves on no-pilot uniform for only one of the five pilot realizations. These results complement the overhead failures in Table 2; both favorable and unfavorable cases are retained. The supplementary CSV reports all 120 fixed-pilot configurations, 2,000 production resamples per configuration and six steps using the same resamples. Figure 2 separates the local risk approximation from its worst-case certificate. Figure 3 retains all original allocation controls, while Figure 4 displays the separate concentration diagnostic.

5.4 Common-context versus independent feedback

An additional exact covariance diagnostic compares independent eight-context samples per expert with eight common contexts observed for all five experts. Both schemes reveal 40 training entries from an empty cache, with no pilots. For each of the 30 real-data training tables, the frame and the 32 training design features are initialized exactly as in the routing benchmark, at $\tau = 0.25$. The finite-population formulas use the full reward table only for this post hoc diagnostic; they select no method, rate or checkpoint.

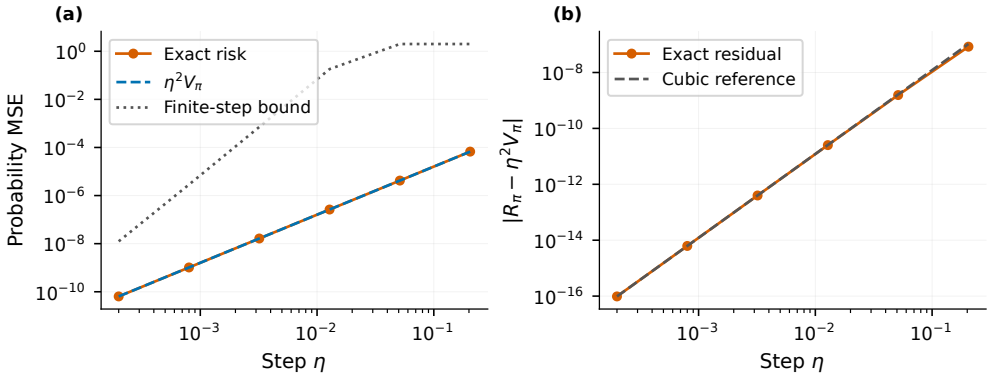

Fig. 2 Exact finite-table calculation over 1000 production-subset combinations. (a) Actual probability mean-square error, its leading quadratic term and the probability-capped finite-step upper bound. (b) Absolute residual after subtracting the leading term, with a cubic reference. The reference shows order, not a fitted statistical confidence band. The bound is valid but conservative.

Table 4 Exact common-context divided by independent covariance trace at the initial state. Entries are the mean [minimum, maximum] over ten split seeds. A ratio below one favors common-context acquisition for this fixed-state variance measure. All-correct is the fraction of training rows on which every expert is correct. These are descriptive ranges, not confidence intervals.

Dataset	Gradient ratio	Policy ratio	All-correct
Digits	0.159 [0.121, 0.188]	0.158 [0.077, 0.222]	0.389
Breast cancer	0.029 [0.020, 0.041]	0.017 [0.012, 0.024]	0.888
Wine	0.035 [0.016, 0.058]	0.031 [0.013, 0.044]	0.858

Proposition 4.3 identifies the difference as a sum of cross-expert covariance terms. Common-context policy variance is smaller in 30/30 matched initial states. When all experts receive reward one on a context, their probability gradients sum to zero, so observing the whole row preserves that cancellation. Independent sampling generally does not. The computed cross terms quantify this coupling on the actual training tables. They do not establish a causal attribution for the final benchmark rankings, whose states, acquisition histories and selected rates evolve.

5.5 Budgeted routing among frozen CPU experts

Three public classification datasets provide a real-data test of the finite-table setting. Every expert and feature scaler is fitted on a disjoint expert-fit partition before router training. Ten prespecified stratified splits are used. These are repeated splits of three fixed datasets, not thirty independent application populations. Feedback is deterministic correctness of a frozen expert; the sampler's randomness comes from context selection. No language model is called.

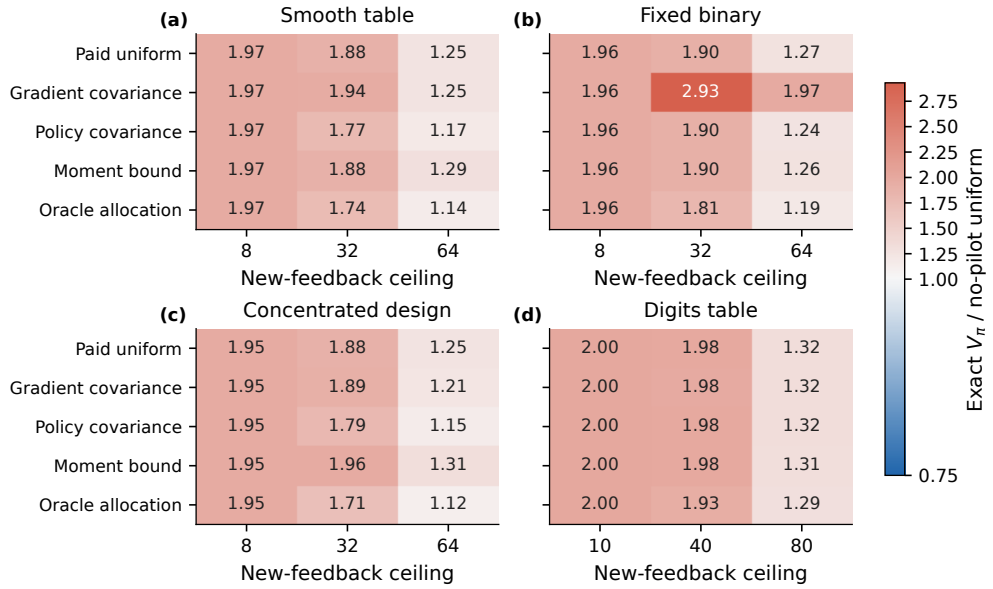

Fig. 3 Allocation versus information overhead in the four original fixtures. Heatmap entries are ratios of exact conditional linearized policy variance to no-pilot uniform sampling at the same new-feedback ceiling. All cells use the specified fixed cache and pilot realization. A ratio below one favors the row method for that fixture; comparisons do not average over future tasks.

The acquisition fractions are $0.1, 0.25, 0.5, 0.8$ of the training response matrix. Each iterative method uses the same seven step candidates and development-only selection. Ridge payoff regression uses seven regularization candidates and the same per-fit response budget, but has a different policy family and deploys a deterministic action. The complete-row control buys all expert outcomes at each chosen context; sub-row credits carry forward, and only the final remainder below one complete row remains unspent.

The reporting slice is fixed for readability. All fractions, penalties, candidates, development scores, selected test values and failure-free budget ledgers are retained in the accompanying CSV files. No test outcome chooses a step, regularization value, method, or checkpoint. Matched-allocation comparisons and method limitations are interpreted with the full results, rather than as a universal superiority claim.

5.6 Interpreting utility and selection

The policy-covariance minus no-pilot uniform mean utility differences at the reporting slice are 0.031, 0.207, -0.436 percentage points for Digits, Breast Cancer and Wine, respectively. These small and mixed differences do not establish a stable utility advantage. Complete-context sampling, payoff regression and development-selected fixed routing remain strong controls (Table 5). The probability-risk criterion is therefore interpreted as a local sampling diagnostic, rather than a sufficient reason to prefer the geometric router.

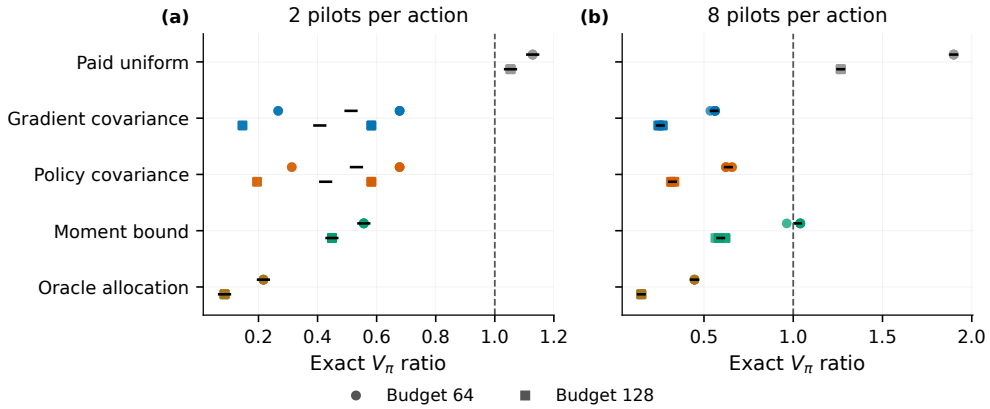

Fig. 4 A deliberately constructed variance-concentration diagnostic added after the initial audit. Dots retain five pilot realizations of one fixed table, with offsets only for readability; horizontal marks show their mean. Ratios are relative to no-pilot uniform sampling. This example illustrates amortization of pilot cost and does not establish an advantage of policy weighting over gradient weighting.

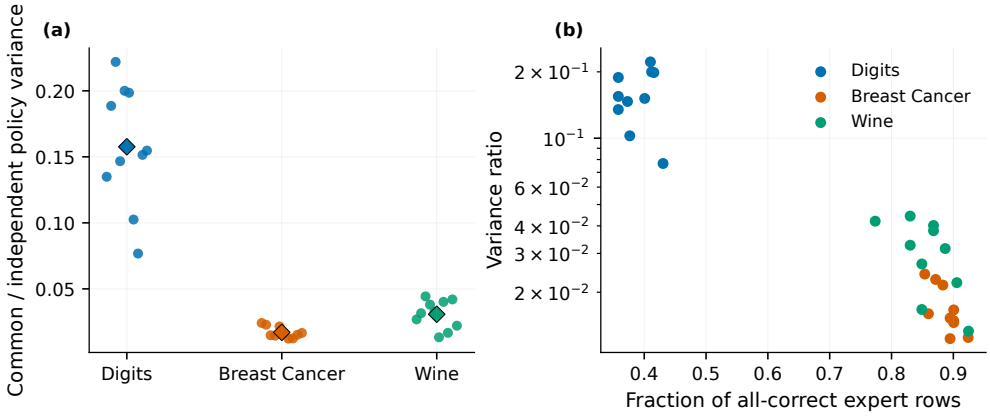

Fig. 5 Exact initial-state covariance diagnostics on the 30 actual expert tables. (a) Matched common-context to independent-stratum policy-variance ratios; dots are ten splits and diamonds their mean. (b) The same ratios against the fraction of training rows on which all experts are correct. These are full-table diagnostics, not information available without acquisition, and they do not establish a causal explanation of final training outcomes.

The highest rate, 256, is selected in 32, 22, 35, 35, 27, 43 of 240 groups for uniform, paid uniform, gradient covariance, policy covariance, upper-bound and complete-context sampling, respectively. The extended grid reduces dependence on the original boundary but does not establish optimal tuning. Ridge also frequently selects an extreme regularization value. These outcomes and the small development sets limit ranking precision; the study makes no confirmatory superiority or equivalence claim.

Table 5 Independent-test expected utility at the prespecified reporting fraction 0.5 and cost penalty 0.1. Costs are a fixed normalized relative-cost scenario, not money, energy or measured latency. Entries are mean (SD) over ten dataset splits; SD is descriptive.

Method	Digits	Breast cancer	Wine
Uniform strata	0.8490 (0.0135)	0.9000 (0.0181)	0.8991 (0.0262)
Uniform + pilots	0.8490 (0.0116)	0.8991 (0.0206)	0.8939 (0.0265)
Gradient covariance	0.8457 (0.0116)	0.8961 (0.0216)	0.9036 (0.0139)
Policy covariance	0.8494 (0.0139)	0.9021 (0.0172)	0.8948 (0.0225)
Policy upper bound	0.8421 (0.0093)	0.8998 (0.0197)	0.8959 (0.0281)
Complete context	0.8654 (0.0088)	0.9120 (0.0155)	0.9097 (0.0240)
Ridge payoff regression	0.9270 (0.0084)	0.9243 (0.0209)	0.9328 (0.0247)
Uniform strategy	0.7869 (0.0155)	0.8978 (0.0179)	0.8933 (0.0154)
Dev-selected fixed	0.9328 (0.0077)	0.9252 (0.0201)	0.9472 (0.0229)

Figures 6 and 7 report held-out outcomes and split-level differences. Figure 8 retains the development-selection and feedback-accounting diagnostics.

6 Discussion

6.1 What the analysis adds

The central object is the probability error of a sampled finite update relative to the full-information update at the same state. The finite-table formulation specifies what is random and what has already been observed. Its exact cache term permits a pilot-dependent production allocation without treating stored values as fresh responses. The policy differential then converts the remaining gradient covariance into a criterion with an observable interpretation.

These steps add a specific risk analysis to established geometry and sampling tools. Stiefel coordinates, canonical gradients, polar updates, finite-population variance and square-root allocation are not claimed as inventions. Table 6 summarizes the distinction.

The upper envelope and integer sensitivity result distinguish two issues that can otherwise be conflated: optimizing an allocation for supplied weights and learning useful weights from limited information. The former can be solved exactly for this unit-cost objective; the latter can remain difficult. Neither the plug-in rule nor the conservative upper weights establish a new general variance-reduction principle.

6.2 When adaptivity pays for its information

The original fixed fixtures show that better allocation among production samples can fail to recover the feedback spent on pilots. The constructed concentration example demonstrates the opposite possibility, while retaining pilot realizations in which the gains are smaller. It is a boundary calculation, not evidence that strong heterogeneity is common in applications. Gradient-covariance allocation also remains competitive and can be better. The numerical conclusion is therefore conditional: useful sensitivity

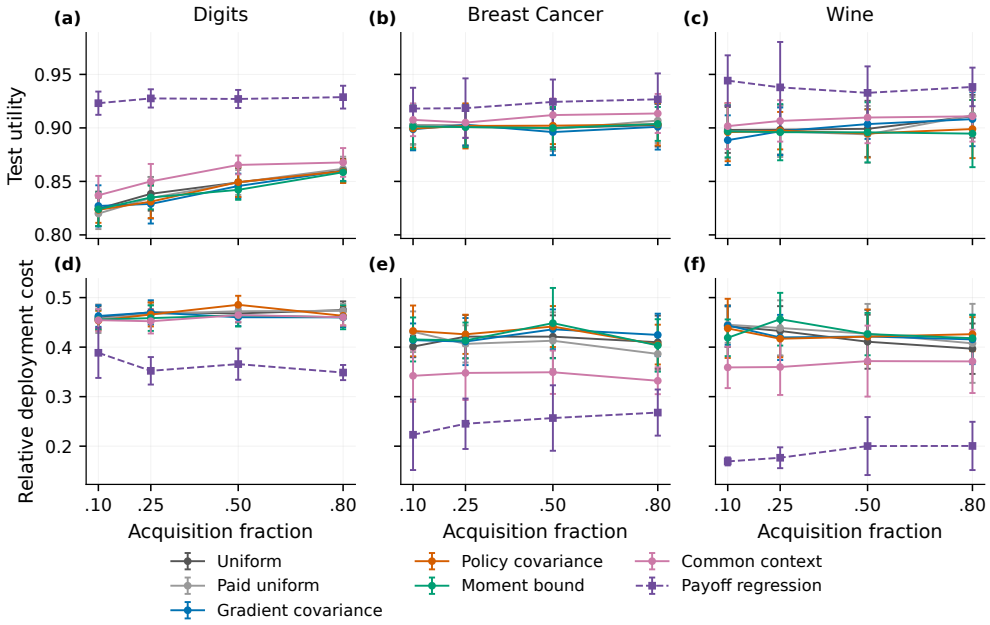

Fig. 6 Independent-test utility (top) and expected relative deployment cost (bottom) at penalty $\lambda = 0.1$. Curves connect the four evaluated acquisition fractions. Symbols show ten-split means; bars show descriptive standard deviations across overlapping splits, not population confidence intervals. The ridge router uses a different policy family and argmax deployment.

weights, sufficient heterogeneity and the cost of estimating them must be considered together.

The complete-context control addresses a different variance mechanism. Independent action strata discard cross-expert covariance. When several experts agree on the same contexts, their probability-gradient contributions can cancel under common-context sampling. A policy-aware allocation over independent strata does not recover that cancellation. Conversely, querying a complete row can be wasteful when only a few experts carry informative variation. This distinction motivates reporting both controls rather than interpreting every sampling improvement as an effect of the chosen metric.

6.3 Routing utility and model capacity

Local policy risk is not a surrogate theorem for final task accuracy. The reference full-gradient step may itself be poorly tuned, and the fixed-temperature orthogonal policy has a restricted logit scale. A small variance relative to that reference does not establish that the policy family is preferable to a deterministic fixed expert or a regression router. The real-data comparisons retain these alternatives and do not establish a practical requirement for Stiefel routing on the evaluated datasets.

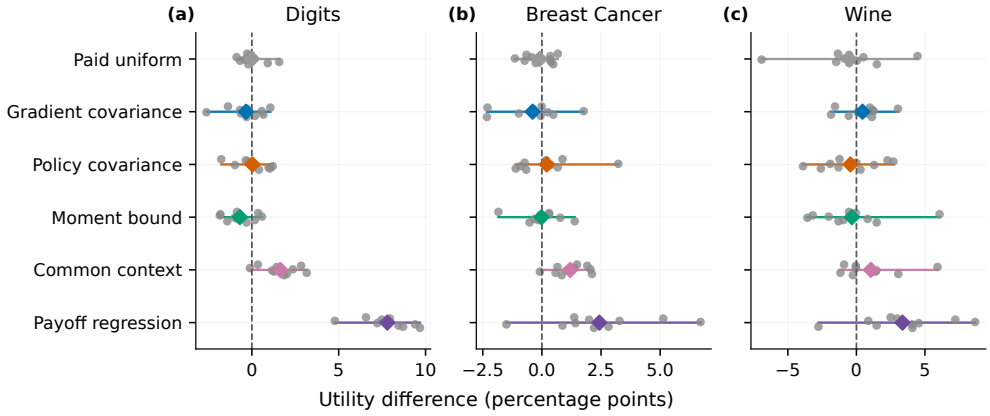

Fig. 7 Paired independent-test utility differences relative to no-pilot uniform routing at acquisition fraction 0.5 and $\lambda = 0.1$. Gray points are ten split-level differences, colored diamonds their mean and segments their observed range. The same source dataset appears in every split; the display is descriptive.

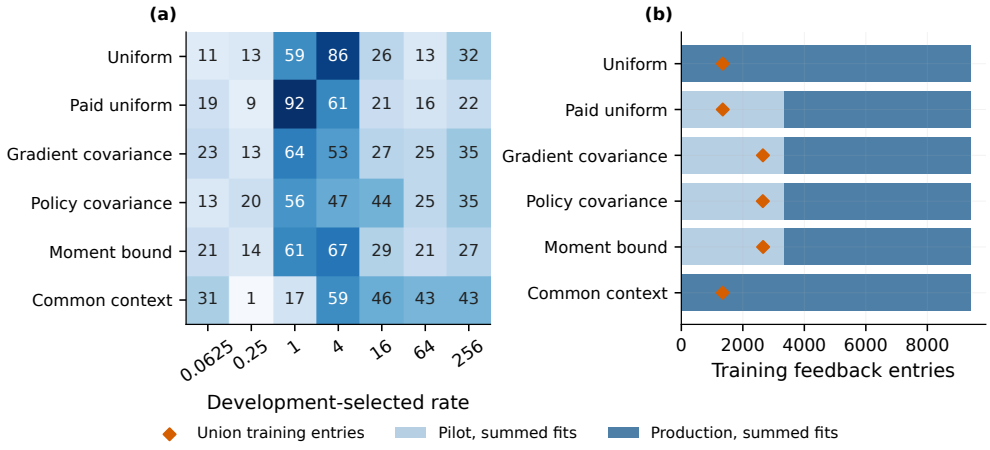

Fig. 8 Selection and resource diagnostics. (a) Development-selected rate counts over all dataset, split, budget and penalty groups for each Stiefel method. (b) Digits, fraction 0.5, $\lambda = 0.1$: mean pilot and production charges summed over seven separately trained candidates, together with the union of revealed training entries. Repeated use of the same feedback across candidates is distinguished from per-fit charges. Development labels and numerical oracle diagnostics are accounted for separately.

This also limits the interpretation of argmax diagnostics. Changing from a stochastic policy to its maximum-probability action changes deployment behavior and can improve or degrade accuracy. It is not the probability-risk quantity proved in Section 4. Likewise, learning temperature, adding unrestricted bias, or changing the expert pool changes the model being compared and should be assessed with explicit capacity controls.

Table 6 Established tools and the additional questions addressed in this paper.

Prior framework	Established object	Specific extension studied here
Stiefel geometry and matched metrics (Edelman et al. 1998; Zheng et al. 2025)	Feasible directions, retractions and representation-dependent algorithms	Sampling risk after a specified finite policy update; compatible realization as a control
Stratified and adaptive sampling (Neyman 1934; Zhao et al. 2025)	Variance weights and budget allocation	History-conditioned cache correction, policy-visible covariance and nonlinear risk remainder
Contextual selection (Li et al. 2026; Pacchiano et al. 2025)	Context-dependent information allocation and constrained decisions	Finite response-table acquisition, with its own feedback and cost definitions
Frozen-model intervention routing (Yin and Zhang 2026)	Choosing interventions using predictive features	Explicit acquisition restrictions and an auditable analysis of the router update

6.4 Limitations and further work

The certificates can be much larger than the observed error. Their value is to state a valid conditional quantity and expose its conservatism, not to suggest that every step should be accepted using the resulting worst-case threshold. The design contexts approximate a chosen feature distribution; they do not supply a population generalization guarantee. The analysis is local in the current state and does not prove stability of independently evolving trajectories, global stationarity at a budget-dependent stopping time, or contextual regret.

The empirical study uses three small public datasets and repeated overlapping splits. Its cost penalty is an explicit relative-cost scenario. The development sets can be small, especially for Wine, and the candidate grids are finite. The grid was expanded using development-boundary diagnostics after an initial exploratory run whose test summaries had already been computed. The final test records are independent of each fitted candidate's training and development selection, but are not a sealed confirmatory evaluation of the entire research process. All initial and final records are retained.

A larger response-table benchmark, measured service costs and a new independent evaluation would strengthen deployment evidence. For a language-model application, repeated generations require genuine additional responses; reusing a public score cannot create them. Unequal acquisition costs and joint action-group sampling also require allocation rules beyond the unit-cost independent-stratum optimization analyzed here.

7 Conclusion

We analyzed budgeted frozen-expert routing through the probability error of a sampled finite update. A cache-corrected estimator yields conditional finite-population covariance, which supports a policy-visible allocation criterion, a nonlinear remainder bound and explicit unknown-moment controls. The experiments verify these relations

and expose the costs of pilot information and ignored cross-expert covariance. They support an auditable criterion for assessing adaptive feedback allocation, rather than a universal claim that a geometric router or a pilot-based rule improves utility.

A Representation compatibility as an implementation control

This appendix relates the contrast implementation to a redundant full Stiefel frame. It does not propose a new manifold or a general retraction. Let $e_0 = \mathbf{1}/\sqrt{m}$, let $C^\top C = I$ and $C^\top e_0 = 0$, and write

$$\Phi = UC^\top + ve_0^\top, \qquad U^\top v = 0, \quad \|v\| = 1.$$

Then $\Phi^\top \Phi = I_m$ and $\Phi C = U$. For every context,

$$\mathrm{softmax}(\Phi^\top h/\tau) = \mathrm{softmax}(CU^\top h/\tau),$$

because the omitted component $(v^\top h)e_0/\tau$ is a common logit shift.

Let G be a fixed ambient direction for the reduced objective, and define

$$g = G - UG^\top U, \qquad b = v^\top G, \qquad \Omega = U^\top G - G^\top U.$$

In column coordinates $[C, e_0]$, the matched full canonical tangent is $Z = [g, -Ub^\top]$ at $X = [U, v]$. Standard canonical geometry gives

$$K(Y, \Delta) = \Delta Y^\top - Y\Delta^\top - Y(Y^\top \Delta)Y^\top, \qquad \mathrm{Exp}^{\mathrm{can}}_Y(\eta\Delta) = e^{\eta K(Y,\Delta)}Y.$$

Direct substitution yields $K(X, Z) = K(U, g)$. Thus the full canonical exponential, after forgetting its last column, equals the reduced canonical exponential for every real η. This is a specialization of classical Stiefel geometry (Edelman et al. 1998); it requires matched directions and updates, consistently with the algorithmic equivalences studied by Zheng et al. (2025).

For the standard polar map,

$$\mathrm{Retr}_Y(\eta\Delta) = (Y + \eta\Delta)(I + \eta^2\Delta^\top\Delta)^{-1/2},$$

expansion at zero instead gives

$$\mathrm{Retr}_X(\eta Z)_{[:,1:m-1]} - \mathrm{Retr}_U(\eta g) = \frac{\eta^2}{2} vb\Omega + O(\eta^3).$$

Indeed, $Z^\top Z$ has top-left block $g^\top g$, bottom-left block $-bU^\top g = -b\Omega$, and multiplication by $[U, v]$ produces $Ug^\top g - vb\Omega$. Subtracting the reduced expansion proves the formula. This is a difference between two selected polar implementations; it is not an unavoidable loss of policy equivalence.

There is also a local compatible completion. For $U_+ = \mathrm{Retr}_U(\eta g)$, set

$$w = (I - U_+ U_+^\top)v, \qquad v_+ = w/\|w\|, \qquad \Phi_+ = U_+ C^\top + v_+ e_0^\top.$$

Where $w \neq 0$, the full frame is feasible and $\Phi_+ C = U_+$ exactly. Its initial tangent is $[g, -Ub^\top]$. The sufficient condition $|\eta| \, \|g\|_F < 1$ ensures $w \neq 0$, since the polar chord bound gives $\|U_+ - U\|_F \leq |\eta| \, \|g\|_F$ and hence $\left\|U_+^\top v\right\| < 1$. The implementation must guard the projected norm when using a completion outside this sufficient region. This construction is a local lift along these directions, not a retraction on every full-frame tangent.

The main analysis therefore works directly with U. Its risk compares sampled and full-information reduced updates, and remains nonzero when an implementation is exactly compatible. Removing a redundant completion does not remove stochastic feedback error or establish superior utility.

B Reproducibility and evidence boundaries

The release separates data construction, mathematical audits, routing experiments and rendering. The main entry points are `src/build_expert_tables.py`, `src/risk_allocation_audit.py`, `src/run_router_benchmark.py`, and `src/plot_paper.py`. The configuration and numerical records are stored under `protocol/`. The test suite checks partition disjointness, feature fitting, the finite-population estimator, covariance identities, integer budgets and development-only selection. Checks of a known mathematical identity are distinct from evidence of a useful router; both are reported with their appropriate interpretation.

Each table manifest records source arrays, split indices, feature transformations, expert settings, software versions and hashes. Random splits are used within the bundled Digits subset and are not claimed to preserve the original collection's writer-independent train/test protocol. The small Wine test partitions in particular limit precision. The repeated-split distribution is not a confidence interval for performance over all future datasets.

The experiments have two different forms of randomness. Production-subset resampling at a fixed finite table measures the estimator's conditional risk. Repeated end-to-end splits additionally change fitted experts, router training records and development selection. They must not be pooled as independent repeated generations of one expert response. A cache entry has the same value on every access.

CPU timing probes use warm-up calls and repeated predictions. Their units and aggregation are recorded in the table manifests. They are affected by local scheduling and batching and are not used to infer a financial or energy saving. The explicit relative deployment-cost scenario is kept separate from these timings.

The public-response-table application has the same finite-action form if its features, outcomes and costs satisfy the stated assumptions. No such language-model table was executed in this study. Extending the results to current model versions, live failures, token limits, or provider-side randomness requires additional data and a revised execution ledger.

References

Absil PA, Malick J (2012) Projection-like retractions on matrix manifolds. SIAM Journal on Optimization 22(1):135–158. https://doi.org/10.1137/100802529

Absil PA, Mahony R, Sepulchre R (2008) Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, URL https://sites.uclouvain.be/absil/amsbook/

Aeberhard S, Forina M (1992) Wine. https://doi.org/10.24432/C5PC7J, URL https://archive.ics.uci.edu/dataset/109/wine

Alpaydin E, Kaynak C (1998) Optical recognition of handwritten digits. https://doi.org/10.24432/C50P49, URL https://archive.ics.uci.edu/dataset/80/optical+recognition+of+handwritten+digits, bundled 1797-sample scikit-learn subset used

Ban Y, Yan Y, Banerjee A, et al (2026) Neural exploitation and exploration of contextual bandits. Journal of Machine Learning Research 27(55):1–38. URL https://jmlr.org/papers/v27/23-0582.html

Boumal N (2023) An Introduction to Optimization on Smooth Manifolds. Cambridge University Press, https://doi.org/10.1017/9781009166164

Carpentier A, Munos R, Antos A (2015) Adaptive strategy for stratified monte carlo sampling. Journal of Machine Learning Research 16(68):2231–2271. URL https://www.jmlr.org/papers/v16/carpentier15a.html

Chadli K, Botterweck G, Saber T (2026) PEARL: Performance and energy aware routing for LLMs. Future Generation Computer Systems 176:108218. https://doi.org/10.1016/j.future.2025.108218

Davis D, Drusvyatskiy D, Shi Z (2025) Stochastic optimization over proximally smooth sets. SIAM Journal on Optimization 35(1):157–179. https://doi.org/10.1137/20M1320225

Edelman A, Arias TA, Smith ST (1998) The geometry of algorithms with orthogonality constraints. SIAM Journal on Matrix Analysis and Applications 20(2):303–353. https://doi.org/10.1137/S0895479895290954

Ha Y, Shashaani S, Pasupathy R (2025) Complexity of zeroth- and first-order stochastic trust-region algorithms. SIAM Journal on Optimization 35(3):2098–2127. https://doi.org/10.1137/24M1664484

Hoeffding W (1963) Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association 58(301):13–30. https://doi.org/10.1080/01621459.1963.10500830

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47(260):663–685. https://doi.org/10.1080/01621459.1952.10483446

Levin E, Kileel J, Boumal N (2025) The effect of smooth parametrizations on nonconvex optimization landscapes. Mathematical Programming 209:63–111. https://doi.org/10.1007/s10107-024-02058-3, published online 4 March 2024

Li D, Chew EP, Li H, et al (2026) Efficient simulation budget allocation for contextual ranking and selection with quadratic models. European Journal of Operational Research 328(3):862–876. https://doi.org/10.1016/j.ejor.2025.08.042

Neyman J (1934) On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society 97(4):558–625. https://doi.org/10.1111/j.2397-2335.1934.tb04184.x

Pacchiano A, Ghavamzadeh M, Bartlett P (2025) Contextual bandits with stagewise constraints. Journal of Machine Learning Research 26(170):1–57. URL https://jmlr.org/papers/v26/24-0267.html

Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12:2825–2830. URL https://jmlr.org/papers/v12/pedregosa11a.html

Wang H, Vandebril R, Van der Veken J, et al (2026) Manifold-valued function approximation from multiple tangent spaces. SIAM Journal on Scientific Computing 48(3):A1734–A1760. https://doi.org/10.1137/25M1753863

Wolberg W, Mangasarian O, Street N, et al (1993) Breast cancer wisconsin (diagnostic). https://doi.org/10.24432/C5DW2B, URL https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

Yin S, Zhang R (2026) Failure-mode-aware uncertainty intervention routing for large language models. Knowledge-Based Systems 351:116685. https://doi.org/10.1016/j.knosys.2026.116685, part B

Zhao B, Wang L, Liu Z, et al (2025) Adaptive client sampling in federated learning via online learning with bandit feedback. Journal of Machine Learning Research 26(8):1–67. URL https://jmlr.org/papers/v26/24-0385.html

Zheng S, Huang W, Vandereycken B, et al (2025) Riemannian optimization using three different metrics for hermitian psd fixed-rank constraints. Computational Optimization and Applications 91:1135–1184. https://doi.org/10.1007/s10589-025-00687-8