FockTrust: Photonic Natural Gradients in Detector-Count Space

Anonymous authors
Paper under double-blind review

Abstract

Photonic learners observe detector counts rather than optical states, so their natural gradient must follow the distribution actually measured. Loss, missed photons, dark counts, and readout alter which parameter directions remain distinguishable. We introduce FockTrust, which defines the Fisher metric directly in detector-count space. Shared base and parameter-shift histograms estimate the task gradient and full metric, while covariance-whitened readout moments map count geometry into tangent space. We also bound the effect of finite counts on the resulting step. In matched 2×2 experiments at $d = 32$, retaining off-diagonal count couplings reduces final loss by 17.6–24.3% relative to diagonal directions under either trust rule. After method-wise tuning on disjoint development cells, full count geometry lowers loss by 25.7% on 12 unseen graphs. The advantage over state-QFIM persists in a two-photon study. FockTrust makes the detector model part of optimization and identifies detector-count space as the operative geometry of measured photonic learning.

1 Introduction

Natural gradient follows the geometry of a probability model (Amari, 1998; Martens, 2020). In photonic learning, the choice of that model is physical. Control parameters prepare a Fock state, but the optimizer sees the state only after photon loss, no-click events, unequal detector efficiencies, dark counts, phase perturbations, and readout. Pure-state quantum natural gradient (QNG) measures distinguishability before this observation process (Stokes et al., 2020), while mixed-state extensions describe noisy or nonunitary preparation through a density-operator quantum Fisher information matrix (QFIM) (Koczor & Benjamin, 2022). Neither geometry directly answers a simpler operational question: which parameter changes can the detector actually distinguish?

This question matters whenever the detector channel is noninvertible. The channel can erase a direction, amplify another, or couple parameters that look separate on the state manifold. A scalar learning rate cannot recover such a change. We therefore define the natural gradient in the categorical law $q_\theta(z \mid x, \xi)$ delivered by the detector. This choice fixes the statistical model first; full, diagonal, block, and compressed matrices then become approximations to the same observed geometry.

FockTrust turns this idea into a practical update. Ordinary base and parameter-shift histograms provide the task gradient, the count-Fisher metric, and a local detector-KL scale, without a second collection devoted to geometry (Figure 1). Because the gradient and metric reuse the same multinomial observations, we analyze their errors jointly and track the effects of finite counts, shrinkage, and exponential averaging. We further show that covariance-whitened readout moments form a positive-semidefinite projection of count Fisher. This result describes exactly which tangent directions survive a compressed readout.

The experiments isolate the mechanism before comparing complete optimizers. A 2×2 intervention at $d = 32$ changes the matrix used for the search direction independently from the matrix used to scale the step. This isolates the contribution of off-diagonal count couplings. We then tune full count geometry, diagonal count geometry, exact-QFIM natural

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.

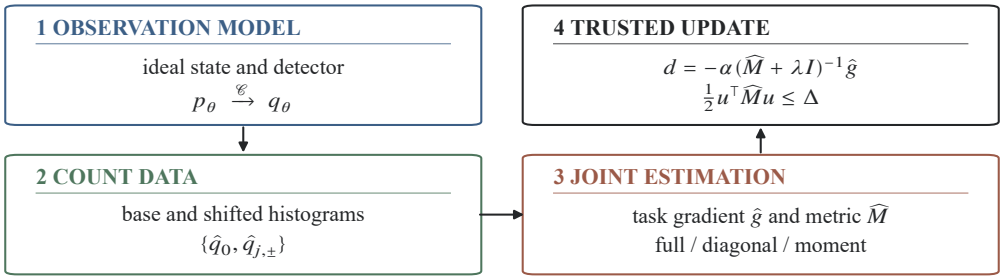


Figure 1: Natural gradient in detector-count space. Base and shifted histograms estimate the task gradient and count-Fisher metric; full, diagonal, and moment forms define alternative updates in the same observed probability space.

gradient, and Adam on equally sized development searches and evaluate them on 12 new graph circuits. Further experiments compare shared counts with disjoint gradient–metric splits and move from one photon to a six-mode, two-photon model.

The paper makes three contributions. First, it places photonic natural gradient in the probability space seen by the detector and characterizes compressed moment geometry as a tangent-space projection of count Fisher. Second, it derives a same-count estimator and a finite-count perturbation analysis that respects the dependence between gradient and metric. Third, the experiments trace the gain to off-diagonal direction geometry, find 25.7% lower final loss than diagonal count geometry on 12 unseen graph circuits, and show how count reuse and photon number change the benefit.

2 Related work

Measurement-defined natural gradients. QNG uses the local geometry of a parameterized quantum state (Stokes et al., 2020); photonic demonstrations and extensions address optical circuits, noisy dynamics, and nonunitary evolution (Yao et al., 2022; Wang et al., 2023; Koczor & Benjamin, 2022). Quotient-space QNG separates state-preserving parameter redundancies from physical soft modes of the state metric (Chen, 2026). A measurement induces a classical Fisher matrix bounded above by the state QFIM (Braunstein & Caves, 1994), providing an experimentally accessible geometry (Meyer, 2021; Lu & Sha, 2025). Random Natural Gradient samples measurement bases and uses their classical Fisher as a low-measurement surrogate for state geometry (Kolotouros & Wallden, 2024); related random-measurement methods accelerate imaginary-time evolution (Kolotouros et al., 2025). FockTrust instead treats the deployed detector channel as the statistical model of the task. Its Fisher describes the distinctions that remain visible after the observation process, rather than reconstructing latent state geometry. Detector quantum Fisher information addresses a complementary problem: it optimizes probes for estimating unknown measurement operators (Das et al., 2026). We hold the detector model fixed and differentiate the observed count law with respect to the photonic model parameters.

Gill and Pal derive loss-aware ambient and conformal geometries whose principal optimizers rescale the effective step while preserving the underlying natural-gradient direction (Gill & Pal, 2026). Detector-count Fisher can also rotate that direction because it changes the probability model on which the gradient is defined.

Readout and compressed information. Readout-Rank Laws derive rank laws for isotropic tangents under computational-basis readout (Ait Haddou, 2026). Our projection result holds for any specified q, J, A and any categorical detector channel. The resulting moment-information matrix is the optimal linear estimating-function information of the empirical feature means. Geometrically, it is a covariance-whitened projection of the count-Fisher

tangent space. This view places full, diagonal, and moment geometry inside one observed statistical model.

Finite-count optimization. Parameter-shift and stochastic-metric methods reduce quantum derivative costs, including exact shift rules for variational photonic circuits (Gacon et al., 2021; Pappalardo et al., 2025; Hoch et al., 2025). Measurement-allocation schemes distribute observations across operators, derivatives, inputs, or iterations (Arrasmith et al., 2020; Gu et al., 2021; Moussa et al., 2023), and metric-aware allocation propagates uncertainty through a regularized natural step (van Straaten & Koczor, 2021). Here, the same categorical observations enter both gradient and geometry. We analyze their dependence and compare reuse with disjoint partitions at the same numbers of emitted shots and circuit evaluations.

Classical preconditioning. K-FAC approximates Fisher curvature through Kronecker factors, whereas Shampoo uses tensor-structured preconditioners (Martens & Grosse, 2015; Gupta et al., 2018). These methods address scalable matrix structure once a preconditioner has been specified. Our plug-in matrix estimates the model Fisher of q_θ, in contrast to the supervised-learning empirical Fisher analyzed by Kunstner et al. (2019).

3 Natural gradient in the detector's probability space

3.1 The distribution seen by the learner

For input x, parameters $\theta \in \mathbb{R}^d$, and detector condition ξ, the observed outcome follows

$$z \sim q_\theta(z \mid x, \xi), \qquad z \in \mathcal{Z}. \tag{1}$$

The distribution q_θ includes detector inefficiency and a no-click outcome; the perturbation study additionally applies dark-count replacement and Gaussian phase jitter. We distinguish this observed law from the ideal circuit law p_θ and from the empirical frequency vector $\widehat{q}_\theta$. The training loss $\mathcal{L}(\theta)$ depends on the observed outcomes or on features computed from them.

3.2 Count Fisher and the directions preserved by readout

At fixed (x, ξ), let $q \in \mathbb{R}^m$ collect the positive outcome probabilities and let $J = \partial q/\partial\theta$. The pullback of categorical Fisher information is

$$F_{\text{count}}(\theta; x, \xi) = J^\top \operatorname{diag}(q)^{-1} J = \sum_{z \in \mathcal{Z}} q_\theta(z \mid x, \xi) \nabla \log q_\theta(z \mid x, \xi) \nabla \log q_\theta(z \mid x, \xi)^\top. \tag{2}$$

For multiple inputs, we average the input-conditioned metrics using the weights of the training objective. This metric is also the quadratic term of local detector KL:

$$\mathrm{KL}(q_\theta \,\|\, q_{\theta+d}) = \tfrac{1}{2} d^\top F_{\text{count}}(\theta) d + o(\|d\|^2). \tag{3}$$

Suppose the readout keeps r features whose values form the rows of $A \in \mathbb{R}^{r \times m}$. Their mean, sensitivity, and single-draw covariance are

$$\mu_A = Aq, \qquad J_A = AJ, \qquad \Sigma_A = A(D - qq^\top)A^\top, \quad D = \operatorname{diag}(q). \tag{4}$$

We call

$$G_A = J_A^\top \Sigma_A^+ J_A \tag{5}$$

the moment-information geometry. It is the covariance-whitened sensitivity of the retained empirical means: equivalently, the optimal linear estimating-function (Godambe) information, or the Fisher information of a local Gaussian moment experiment with fixed covariance (Godambe, 1960). In general it is not the exact pushforward Fisher information of the full random variable $A(Z)$. Writing $X = D^{-1/2}J$, whitening the categorical tangent space gives $G_A = X^\top P_A X$ and $F_{\text{count}} = X^\top X$, where P_A is the orthogonal projector onto the

covariance-supported span of the centered features (Appendix A; cf. the isotropic rank-law setting of Ait Haddou, 2026). Consequently,

$$0 \preceq G_A \preceq F_{\text{count}}. \tag{6}$$

One-hot features recover the count Fisher. A compressed readout preserves it exactly when $P_A X = X$, that is, when its centered feature span contains every categorical score direction. The inequality therefore measures how much of the count-Fisher tangent space remains visible to empirical moment means. It is a local projection statement about those means, rather than a global statement about the information in the full feature-valued observation.

3.3 Estimating gradient and geometry from the same counts

Each parameter-shift update samples one base circuit and two shifted circuits per parameter. The resulting empirical distributions are used jointly:

$$\{\widehat{q}_0, \widehat{q}_{j,+}, \widehat{q}_{j,-}\}_{j=1}^{d} \longmapsto (\widehat{g}_t, \widehat{F}_t). \tag{7}$$

Thus metric estimation does not require a second physical sample stream. A fixed shrinkage rule prevents rare empirical outcomes from producing singular inverse probabilities. Metric estimates can be smoothed across updates by an exponential-moving-average (EMA), while multiple inputs are averaged within an update.

More explicitly, for the equally weighted input set $\mathcal{X}$, let $c_{x,a,z}$ be the count of outcome z in arm a at input x, let $N_{x,a} = \sum_z c_{x,a,z}$, and set $f_{x,a} = c_{x,a}/N_{x,a}$. We use a fixed shrinkage fraction $0 < \gamma < 1$,

$$\widetilde{q}_{x,a} = (1-\gamma) f_{x,a} + \frac{\gamma}{m}\mathbf{1}. \tag{8}$$

This keeps the inverse-probability support floor γ/m invariant to arm allocation. If parameter j admits a shift pair (s_j, r_j), its empirical probability Jacobian is

$$\widehat{J}_{x,:,j} = r_j\big(\widetilde{q}_{x,j,+} - \widetilde{q}_{x,j,-}\big), \qquad \widehat{F}_t = \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \widehat{J}_x^{\top}\,\mathrm{diag}(\widetilde{q}_{x,0})^{-1}\widehat{J}_x. \tag{9}$$

We fix γ and every parameter-shift rule before evaluation; Appendix B reports task-specific values and the unequal-arm estimator. Equation (9) is positive semidefinite by construction.

The same ingredients give any probability-smooth task gradient of the form $g(q, J) = J^{\top} b(q)$. For a target law $q_x^{\star}$, categorical KL uses $b(q) = -q_x^{\star}/q$ (componentwise division); moment MSE and linear-residual objectives use their corresponding bounded local residual maps. Appendix A gives the general bounded–Lipschitz perturbation bound and states the scope of the Helmholtz specialization.

For EMA stabilization, $\widehat{M}_1 = \widehat{F}_1$ and $\widehat{M}_t = \beta_t \widehat{M}_{t-1} + (1-\beta_t)\widehat{F}_t$, $0 \le \beta_t < 1$. Experiments use constant $\beta_t = \beta$ (including $\beta = 0$). Thus $\widehat{F}_t$ chooses the detector-facing geometry, while β trades sampling variation against lag; the ablations separate them.

3.4 How count noise perturbs the natural step

Reusing observations makes $\widehat{g}_t$ and $\widehat{F}_t$ dependent. Let $\bar{F}_s$ be the population version of the fixed-shrinkage estimator and $\bar{M}_t$ its matching EMA. A single Hoeffding event per update jointly gives $\|\widehat{F}_s - \bar{F}_s\|_2 \le E_{F,s}$ and $\|\widehat{g}_s - \bar{g}_s\|_2 \le E_{g,s}$. For the exact variable-decay EMA weights $w_{t,s}$, set $E_{M,t} = \sum_{s\le t} w_{t,s} E_{F,s}$. Appendix A states every constant, weight, and multi-input form.

Proposition 1 (EMA same-count step bound). For damping $\lambda > 0$, let $\widehat{v}_t = -(\widehat{M}_t + \lambda I)^{-1}\widehat{g}_t$ and $\bar{v}_t = -(\bar{M}_t + \lambda I)^{-1}\bar{g}_t$. On one joint event of probability at least $1-\delta$,

$$\|\widehat{M}_t - \bar{M}_t\|_2 \le E_{M,t}, \qquad \|\widehat{v}_t - \bar{v}_t\|_2 \le \frac{E_{g,t}}{\lambda} + \frac{E_{M,t}\|\bar{g}_t\|_2}{\lambda^2}. \tag{10}$$

No gradient–metric independence is assumed; $\beta_t = 0$ recovers the single-update bound.

Equation (10) controls sampling error around the population-shrunken EMA step. The implemented update then distinguishes a trusted proposal from the displacement that is ultimately applied. The direction matrix $\widehat{M}_t^{\mathrm{dir}}$ is either the full EMA or its diagonal extraction. Let $\mathcal{S}_t$ be the identity for full trust and diagonal extraction for diagonal trust, with $\widehat{T}_t = \mathcal{S}_t(\widehat{M}_t^{\mathrm{full}})$. Then

$$d_t = -\alpha_t(\widehat{M}_t^{\mathrm{dir}} + \lambda I)^{-1}\widehat{g}_t, \qquad s_t = \min\left\{1, \sqrt{\frac{\Delta_t}{\frac{1}{2}d_t^\top \widehat{T}_t d_t}}\right\},$$
$$c_t = \min\{1, R/\|s_t d_t\|_2\}, \qquad \widetilde{u}_t = c_t s_t d_t. \tag{11}$$

We set a factor to one when its denominator vanishes. In the high-dimensional and two-photon panels, $\mathcal{C} = [0.08, 1.45]^d$ and the actual displacement is

$$a_t = \Pi_{\mathcal{C}}(\theta_t + \widetilde{u}_t) - \theta_t, \qquad \theta_{t+1} = \theta_t + a_t. \tag{12}$$

The scalar contractions guarantee $\frac{1}{2}\widetilde{u}_t^\top \widehat{T}_t \widetilde{u}_t \le \Delta_t$ and $\|\widetilde{u}_t\|_2 \le R$. Euclidean box projection also preserves the norm cap, but need not contract a full-matrix trust quadratic. This distinction matters for the scope of the KL guarantee below. These panels use $R = 0.35$; the earlier two-parameter controls apply the capped proposal directly with $R = 0.5$. FockTrust uses the full observed EMA for direction and trust; the factorial diagonalizes either matrix independently.

Define

$$\Gamma_t = E_{M,t} + L_{\mathrm{EMA},t} + B_{\mathrm{geom},t} + B_{\gamma,t}, \qquad \widehat{\delta}_t(w) = \tfrac{1}{2}w^\top \widehat{T}_t w, \tag{13}$$

where F_t is the current unshrunken detector Fisher, $\bar{F}_t$ its population fixed-shrinkage counterpart, and $\bar{M}_t^{\mathrm{full}}$ the matching full population EMA. Further, $L_{\mathrm{EMA},t} = \|\bar{M}_t^{\mathrm{full}} - \bar{F}_t\|_2$, $B_{\mathrm{geom},t} = \|\mathcal{S}_t(\bar{F}_t) - \bar{F}_t\|_2$, and $B_{\gamma,t} = \|\bar{F}_t - F_t\|_2$. Let $f_t(w) = \mathrm{KL}(q_{\theta_t} \,\|\, q_{\theta_t + w})$. Assume positive support and $f_t \in C^3$ on the R-ball, with $\sup_{\|u\|_2=1,\, \|w\|_2 \le R} |D^3 f_t(w)[u,u,u]| \le L_{\mathrm{KL},t}$. Then, on the same concentration event, every $\|w\|_2 \le R$ satisfies

$$\mathrm{KL}(q_{\theta_t} \,\|\, q_{\theta_t + w}) \le \widehat{\delta}_t(w) + \tfrac{1}{2}\Gamma_t \|w\|_2^2 + \tfrac{1}{6}L_{\mathrm{KL},t}\|w\|_2^3. \tag{14}$$

Consequently, the pre-box proposal obeys the a-priori corollary

$$\mathrm{KL}(q_{\theta_t} \,\|\, q_{\theta_t + \widetilde{u}_t}) \le \Delta_t + \tfrac{1}{2}\Gamma_t R^2 + \tfrac{1}{6}L_{\mathrm{KL},t}R^3. \tag{15}$$

The applied displacement obeys Eq. (14) with $\widehat{\delta}_t(a_t)$; for diagonal trust, coordinatewise projection also preserves the Δ_t corollary. Appendix A proves these claims, quantifies the full-metric projection penalty, and bounds EMA lag by the weighted parameter path.

3.5 Contracting the trust radius near the sampling floor

A constant positive radius can continue to admit nonzero metric steps after the iterate reaches the sampling-noise floor. We instead contract the radius over the last quarter of training. For T updates,

$$\Delta_t = \begin{cases} \Delta_0, & t/T \le \rho, \\ \Delta_0\left[1 - \dfrac{t/T - \rho}{1 - \rho}(1 - \eta)\right], & t/T > \rho, \end{cases} \qquad \rho = 0.75, \quad \eta = 0.10. \tag{16}$$

Here ρ is the onset fraction and η is the terminal radius relative to Δ_0. The schedule depends only on update progress. It uses neither exact validation loss nor additional detector samples.

Cost of one update. With m outcomes and a two-point shift rule, one update uses $1 + 2d$ circuit arms and $(1 + 2d)n$ emitted shots per input. Constructing $\widehat{J}$ costs $O(md)$, forming a dense count Fisher costs $O(md^2)$, and a dense solve costs $O(d^3)$ with $O(d^2)$ memory. Diagonal geometry reduces metric formation and the solve to $O(md)$ and $O(d)$, respectively. A generalized rule with K nonzero shifts per parameter uses $1 + Kd$ arms but leaves the classical metric cost unchanged. Our two-photon experiment has $K = 8$, $d = 8$, and hence 65 arms per update. These costs identify the full matrix, rather than detector samples, as the scaling bottleneck once d becomes large.

Matched physical cost. Methods are compared at equal emitted shots and circuit evaluations, with shared counts charged once. All methods receive the exact target law as part of the synthetic task. All count-geometry variants and Adam otherwise use sampled current-model detector quantities online; the exact-QFIM comparator additionally receives an uncharged exact state metric. Exact current-model detector probabilities and losses are reserved for offline selection, reporting, and calibration.

4 Experiments

Photonic graph family. The principal model contains one photon in $d + 1$ modes and d trainable Givens rotations. Chain, star, broom, and seeded recursive graphs determine which modes share rotations and therefore which parameter scores correlate after detection. Direct number readout retains no-click and one-click outcomes. Four detector profiles have common mean efficiency 0.72 and anisotropy half-range 0.24 but different mode assignments. For each graph and profile, an interior parameter vector defines an in-family target law $q_{\theta^\star}$; all methods minimize $\mathrm{KL}(q_{\theta^\star} \,\|\, q_\theta)$ from paired random initializations. At $d = 32$, every method receives 64 emitted shots for one base and 64 shifted arms over 96 updates: 399,360 shots and 6,240 circuit evaluations per run.

Separating direction geometry from step control. The mechanism experiment crosses four graphs, four detector profiles, four targets, and eight sampling seeds with FF, FD, DF, and DD. The first letter denotes full or diagonal direction geometry, and the second denotes full or diagonal trust geometry. Scalar hyperparameters, emitted shots, and circuit evaluations are matched. Comparing FF with DF and FD with DD measures the effect of direction geometry; the remaining contrasts measure trust geometry and their interaction.

We next compare FF, DD, an exact-QFIM natural-gradient baseline, and Adam after method-specific tuning. Each method receives 12 candidates and the same 96 graph–profile–target–design cells. FF and DD share a grid over learning rate, damping, trust radius, EMA, and terminal contraction. Development loss selects one setting per method, with stability breaking ties. We then evaluate those settings on 12 new graphs, four new targets, and 12 new design seeds under four detector profiles. Graph, target, design, and randomized-profile seeds do not overlap with development. The QFIM optimizer uses its exact state metric and a different step controller, so this comparison asks which complete procedure works better. The factorial above supplies the mechanism test.

Reusing each detected count. To isolate sample reuse, each arm still emits the same 64 shots, but its counts are partitioned without replacement between gradient and Fisher estimators at fractions 25:75, 50:50, and 75:25. Reuse supplies all counts to both estimators. The comparison keeps circuit arms, shots, updates, and optimizer settings identical on Born fitting, high-frequency moment fitting, and a $d = 32$ graph task. A separate calibration crosses $d \in \{16, 32\}$, two detector profiles, full and diagonal geometry, three shot levels, and three stages of optimization. It measures Jacobian, Fisher/EMA, gradient, and natural-step errors, and compares the Fisher quadratic with exact detector KL for pre-box proposals.

A two-photon test. The cross-sector simulation propagates two photons through six passive modes with eight trainable rotations. Lossy number-resolving readout has 28 zero-, one-, and two-photon outcomes and a fixed 0.002 symmetric readout background. For each parameter the outcome probabilities are trigonometric polynomials of degree at most four. An exact nine-node Fourier differentiation rule has eight nonzero shifted nodes, giving 65 shared count arms per update. Mesh and hub circuits, multiple detector profiles, and disjoint development/holdout targets and sampling seeds are used. FF, DD, exact-QFIM natural gradient, and Adam receive four candidates each in development and 64 updates of 64 shots per arm in the 1,152-run holdout. No hardware execution is claimed.

Uncertainty across tasks. All endpoint ratios are computed in the log domain and balance targets, profiles, and graphs. Crossed bootstrap draws resample the four targets and paired design seeds while retaining the evaluated graphs and detector profiles. The intervals

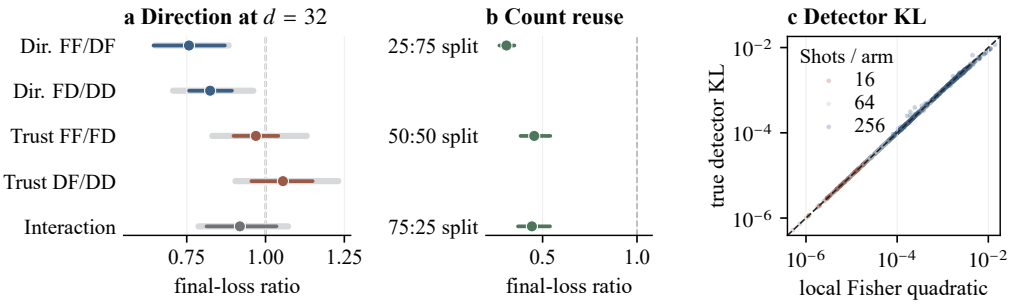


Figure 2: Off-diagonal direction geometry drives the gain. a, Direction and trust ablations at $d = 32$. b, Shared versus split count estimators. c, Fisher quadratic versus detector KL. Dark intervals are marginal 95% CIs; pale intervals are simultaneous.

Table 1: The $d = 32$ factorial favors off-diagonal count directions. Ratios are target-balanced geometric means of final exact loss; values below one favor the first method. U_{95}^{sim} is the simultaneous one-sided 95% upper limit within each prespecified comparison family. Emitted shots and circuit evaluations are matched within every row; exact-QFIM rows receive their metric as an uncharged simulator oracle.

Study	Controlled contrast	Ratio [95% CI]	U_{95}^{sim}
$d = 32$ mechanism	Full/diag. direction, full trust	0.757 [0.645, 0.871]	0.857
$d = 32$ mechanism	Full/diag. direction, diag. trust	0.824 [0.758, 0.894]	0.933
Graph holdout	Full/diagonal count	0.743 [0.702, 0.782]	0.806
Graph holdout	Full count/exact-QFIM NG	0.745 [0.685, 0.805]	0.808
Graph reuse	Reuse/disjoint 25:75	0.308 [0.271, 0.351]	0.834
Two-photon holdout	Full count/exact-QFIM NG	0.653 [0.471, 0.882]	0.843
Two-photon holdout	Full count/Adam	1.278 [1.057, 1.521]	1.651

therefore describe this graph–detector suite rather than a hypothetical population of graph families. Reported intervals use 50,000 synchronized draws. Familywise two-sided intervals use the maximum absolute centered log contrast, while simultaneous one-sided upper limits use the maximum upper centered log contrast within each predefined family. Appendix B gives the grids, seed partitions, per-graph intervals, and complete counts of shots and circuit evaluations.

5 Results

5.1 Off-diagonal count couplings improve the search direction

The factorial asks whether the full matrix helps by changing the search direction or by scaling the step (Figure 2a). Under full trust, replacing the diagonal direction with the full direction gives a final-loss ratio of 0.757 [0.645, 0.871]. Under diagonal trust, the corresponding FD/DD ratio is 0.824 [0.758, 0.894]. Their simultaneous upper limits are 0.857 and 0.933. In comparison, the trust-only ratios are 0.969 and 1.055, with intervals spanning one. The advantage therefore follows the off-diagonal couplings in the search direction and appears under either step controller.

5.2 The advantage carries to unseen graph circuits

After separate development tuning, the full-to-diagonal ratio on 12 unseen graphs is 0.743 [0.702, 0.782], a 25.7% reduction in final loss. Ratios against exact-QFIM natural gradient and Adam are 0.745 and 0.479, with simultaneous upper limits 0.808 and 0.519. Thus the

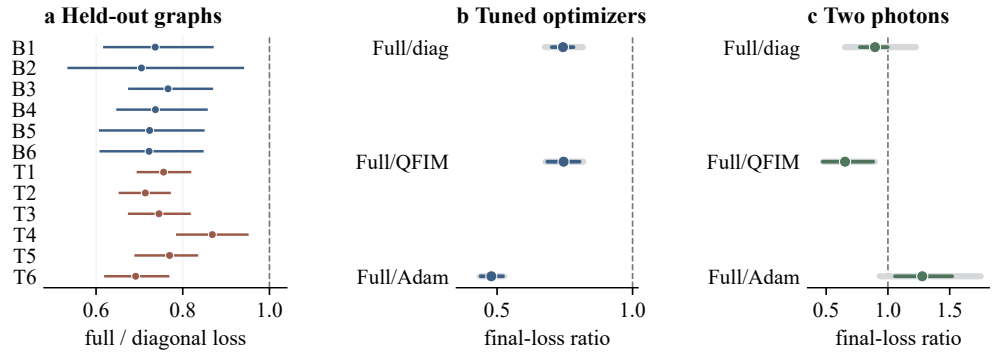


Figure 3: Count geometry transfers to unseen graphs. a, Full-to-diagonal ratios on 12 unseen graphs (B1–B6: broom graphs; T1–T6: recursive trees, in the order of Appendix Table 4). b, Pooled optimizer comparisons. c, Two-photon optimizer comparison. Intervals are 95%; pale intervals are simultaneous.

ordering persists after each optimizer selects its own schedule. The factorial localizes the gain to direction geometry; the QFIM comparison changes both direction and control.

Every graph favors full over diagonal count geometry, and every marginal 95% interval lies below one (Figure 3a). Eight of the 12 intervals remain below one after familywise adjustment, spanning both graph families and all four detector assignments. Appendix Table 4 reports every graph.

5.3 Count reuse and finite-count controls

On the $d = 32$ graph task, count reuse relative to 25:75, 50:50, and 75:25 splits gives final-loss ratios 0.308 [0.271, 0.351], 0.456 [0.383, 0.541], and 0.443 [0.370, 0.539] (Figure 2b), under identical shots, updates, and circuit evaluations. The 25:75 split raises median gradient error from 1.88 to 15.1 and natural-step error from 0.55 to 5.85; its loss ratio remains below one after joint adjustment over nine reuse contrasts ($U_{95}^{\mathrm{sim}} = 0.834$). Low-dimensional controls show task dependence: reuse is strong for high-frequency fitting, whereas Born-fitting comparisons are mixed (Appendix Table 5).

Additional finite-count controls identify the role of probability support and trust contraction. In two low-dimensional controls, removing shrinkage raises median loss by 39.6–495.8, while giving the metric a separate count stream raises it by 14.4–98.0 relative to shared counts. In one high-frequency control, late-stage trust contraction reduces the endpoint by 63.7%. Together they locate the gain in probability support, sample sharing, and the sampling floor. Across seven detector conditions, the FockTrust-to-exact-QFIM median-loss ratio spans 1.5–3.3%; Appendix Tables 9–12 report the complete matrix.

5.4 Count Fisher predicts the local detector KL

Across dimensions, detector profiles, shot levels, geometries, and stored optimization states, the exact-KL/Fisher-quadratic ratio for pre-box proposals has median 0.999 and a 5–95% range of 0.931–1.023. EMA lowers median metric error from 3.63 to 2.39, with path-dependent lag. Analytical bounds cover all measured errors but are loose. The KL agreement concerns the trusted proposal before projection; the projected step obeys the general transfer inequality (Appendix Table 6).

5.5 Two photons change the optimizer ranking

The tuned two-photon holdout gives FF/DD = 0.895 [0.772, 0.998], with both topology point estimates below one but $U_{95}^{\mathrm{sim}} = 1.157$. Full count geometry versus exact-QFIM natural gradient gives 0.653 [0.471, 0.882], with $U_{95}^{\mathrm{sim}} = 0.843$. Thus the full-count advantage over the state-QFIM optimizer extends beyond the single-photon family, although FF/DD is less certain. Adam has the lowest endpoint: FF/Adam = 1.278 [1.057, 1.521], with familywise interval [0.933, 1.751]. Detector geometry is useful without determining the best optimizer.

6 Discussion

The $d = 32$ factorial identifies the search direction as the source of the full-count gain: off-diagonal couplings improve final loss under either trust rule. The effect across 12 held-out graphs links this gain to detector-visible correlations throughout the evaluated graph–detector suite.

On the graph task, count reuse improves final loss against all three disjoint splits at matched physical cost; the Born controls show that the gain depends on the task. EMA lowers median metric error but can lag behind the current geometry. Count Fisher describes local detector KL for pre-box proposals, and the transfer inequality covers projected steps. Diagonal and block approximations can discard detector-visible couplings.

In the two-photon study, full count geometry improves on exact-QFIM natural gradient, whereas the FF/DD adjusted comparison is inconclusive. Adam has the lowest point estimate, with an adjusted interval that includes one for its comparison with full count. Detector geometry and step control remain separate optimizer choices.

Scope and next tests. The experiments cover simulated one- and two-photon detector-KL fitting through $d = 32$. They vary targets and paired streams over the tested graphs and detector profiles, but not model mismatch, drift, or hardware.

Synthetic targets are exact: selection uses simulator loss, updates use sampled detector quantities, and QFIM receives the exact state metric. Hardware tests should match detector trust for a QFIM direction to isolate geometry.

Theory assumes positive outcome probabilities, smooth gradients, and a bounded KL remainder; projection uses the transfer inequality. Dense geometry costs $O(d^2)$ memory and an $O(d^3)$ solve, motivating structured solvers.

7 Conclusion

A photonic learner acts in its detector's probability space. FockTrust learns gradients and full count geometry from shared histograms. Off-diagonal couplings explain the $d = 32$ gain, and tuned full count geometry reduces final loss by 25.7% on 12 unseen graph circuits. In two-photon simulation, full count geometry also outperforms the exact-QFIM comparator.

The comparison with the state QFIM and Adam sharpens this interpretation. Count geometry improves the natural-gradient direction when detector-induced correlations matter, while Adam's lower point estimate in the two-photon study shows that geometry alone does not determine the best step controller. The method's benefit depends on both detector-defined geometry and the rule used to convert it into an update. Because the gradient and metric use the same histograms, loss, missed photons, dark counts, and readout enter at the observed distribution rather than through an idealized state. Adaptive measurement allocation and structured count-metric solvers are the next tests for larger photonic models. Such tests should report detector KL, final loss, shot and circuit cost, and trust-rule selection on disjoint data. These measurements would show whether cheaper structured metrics preserve the observed advantage as the count alphabet and parameter dimension grow.

Reproducibility. Appendices A–B give derivations and experimental details; the release includes code, settings, and data needed to reconstruct the tables and figures.

AI use statement. OpenAI Codex was used for coding assistance and language editing. The authors reviewed all AI-assisted content and take full responsibility for the manuscript, implementation, and reported results.

References

Marwan Ait Haddou. Readout-rank laws for isotropic quantum tangents, 2026. URL https://arxiv.org/abs/2608.07628.

Shun-ichi Amari. Natural gradient works efficiently in learning. Neural Computation, 10 (2):251–276, 1998. doi: 10.1162/089976698300017746.

Andrew Arrasmith, Lukasz Cincio, Rolando D. Somma, and Patrick J. Coles. Operator sampling for shot-frugal optimization in variational algorithms, 2020. URL https://arxiv.org/abs/2004.06252.

Samuel L. Braunstein and Carlton M. Caves. Statistical distance and the geometry of quantum states. Physical Review Letters, 72(22):3439–3443, 1994. doi: 10.1103/PhysRevLett.72.3439.

Zeyu Chen. Quantum natural gradient on quotient spaces, 2026. URL https://arxiv.org/abs/2608.29350.

Aritra Das, Simon K. Yung, Lorcán O. Conlon, Özlem Erkılıç, Angus Walsh, Yong-Su Kim, Ping K. Lam, Syed M. Assad, and Jie Zhao. Precision bounds for characterising quantum measurements. Nature Communications, 17:1821, 2026. doi: 10.1038/s41467-026-68529-7. URL https://doi.org/10.1038/s41467-026-68529-7.

Julien Gacon, Christa Zoufal, Giuseppe Carleo, and Stefan Woerner. Simultaneous perturbation stochastic approximation of the quantum Fisher information. Quantum, 5:567, 2021. doi: 10.22331/q-2021-10-20-567.

Ankit Gill and Kunal Pal. Loss-aware state space geometry for quantum variational algorithms, 2026. URL https://arxiv.org/abs/2604.05627.

V. P. Godambe. An optimum property of regular maximum likelihood estimation. The Annals of Mathematical Statistics, 31(4):1208–1211, 1960. doi: 10.1214/aoms/1177705693.

Andi Gu, Angus Lowe, Pavel A. Dub, Patrick J. Coles, and Andrew Arrasmith. Adaptive shot allocation for fast convergence in variational quantum algorithms, 2021. URL https://arxiv.org/abs/2108.10434.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 1842–1850, 2018. URL https://proceedings.mlr.press/v80/gupta18a.html.

Francesco Hoch, Giovanni Rodari, Taira Giordani, Paul Perret, Nicolò Spagnolo, Gonzalo Carvacho, Ciro Pentangelo, Simone Piacentini, Andrea Crespi, Francesco Ceccarelli, Roberto Osellame, and Fabio Sciarrino. Variational approach to photonic quantum circuits via the parameter shift rule. Physical Review Research, 7(2):023227, 2025. doi: 10.1103/PhysRevResearch.7.023227.

Bálint Koczor and Simon C. Benjamin. Quantum natural gradient generalized to noisy and nonunitary circuits. Physical Review A, 106(6):062416, 2022. doi: 10.1103/PhysRevA.106.062416.

Ioannis Kolotouros and Petros Wallden. Random natural gradient. Quantum, 8:1503, 2024. doi: 10.22331/q-2024-10-22-1503.

Ioannis Kolotouros, David Joseph, and Anand Kumar Narayanan. Accelerating quantum imaginary-time evolution with random measurements. Physical Review A, 111(1):012424, 2025. doi: 10.1103/PhysRevA.111.012424.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical Fisher approximation for natural gradient descent. In Advances in Neural Information Processing Systems, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/46a558d97954d0692411c861cf78ef79-Abstract.html.

Jianfeng Lu and Kecen Sha. Quantum Fisher information matrix via its classical counterpart from random measurements, 2025. URL https://arxiv.org/abs/2509.08196.

James Martens. New insights and perspectives on the natural gradient method. Journal of Machine Learning Research, 21(146):1–76, 2020. URL https://jmlr.org/papers/v21/17-678.html.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pp. 2408–2417, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

Johannes Jakob Meyer. Fisher information in noisy intermediate-scale quantum applications. Quantum, 5:539, 2021. doi: 10.22331/q-2021-09-09-539.

Charles Moussa, Max Hunter Gordon, Michal Baczyk, M. Cerezo, Lukasz Cincio, and Patrick J. Coles. Resource frugal optimizer for quantum machine learning. Quantum Science and Technology, 8(4):045019, 2023. doi: 10.1088/2058-9565/acef55.

Axel Pappalardo, Pierre-Emmanuel Emeriau, Giovanni de Felice, Brian Ventura, Hugo Jaunin, Richie Yeung, Bob Coecke, and Shane Mansfield. Photonic parameter-shift rule: Enabling gradient computation for photonic quantum computers. Physical Review A, 111 (3):032429, 2025. doi: 10.1103/PhysRevA.111.032429.

James Stokes, Josh Izaac, Nathan Killoran, and Giuseppe Carleo. Quantum natural gradient. Quantum, 4:269, 2020. doi: 10.22331/q-2020-05-25-269.

Barnaby van Straaten and Bálint Koczor. Measurement cost of metric-aware variational quantum algorithms. PRX Quantum, 2(3):030324, 2021. doi: 10.1103/PRXQuantum.2.030324.

Yizhi Wang, Shichuan Xue, Yaxuan Wang, Jiangfang Ding, Weixu Shi, Dongyang Wang, Yong Liu, Yingwen Liu, Xiang Fu, Guangyao Huang, Anqi Huang, Mingtang Deng, and Junjie Wu. Experimental quantum natural gradient optimization in photonics. Optics Letters, 48(14):3745–3748, 2023. doi: 10.1364/OL.494560.

Yuan Yao, Pierre Cussenot, Richard A. Wolf, and Filippo Miatto. Complex natural gradient optimization for optical quantum circuit design. Physical Review A, 105(5):052402, 2022. doi: 10.1103/PhysRevA.105.052402.

A Detector-count geometry: complete derivations

This appendix establishes the local statements used in Section 3. The conclusions concern the detector-facing categorical model and the computed update; they do not assert global convergence or a hardware advantage.

A.1 Detector-count and moment geometry

Fix an input and detector condition. Let $q(\theta) = (q_1(\theta), \dots, q_m(\theta))^\top$ be twice continuously differentiable in a neighborhood of θ, with $q_i(\theta) > 0$ and $\mathbf{1}^\top q(\theta) = 1$. Write $D = \mathrm{diag}(q)$ and $J = \partial q / \partial \theta$. Normalization implies $\mathbf{1}^\top J = 0$.

Proposition 2 (Fisher quadratic for detector KL). For a displacement $d \to 0$ that preserves the local support,

$$\mathrm{KL}(q_\theta \,\|\, q_{\theta+d}) = \frac{1}{2} d^\top J^\top D^{-1} J d + o(\|d\|^2). \tag{17}$$

Thus the Hessian of the directed local detector KL is $F_{\mathrm{count}} = J^\top D^{-1} J$.

Proof. Let $\Delta q = q(\theta + d) - q(\theta) = Jd + O(\|d\|^2)$. Componentwise Taylor expansion gives

$$\mathrm{KL}(q \,\|\, q + \Delta q) = -\sum_i q_i \log\left(1 + \frac{\Delta q_i}{q_i}\right) \tag{18}$$

$$= -\sum_i \Delta q_i + \frac{1}{2} \sum_i \frac{\Delta q_i^2}{q_i} + o(\|\Delta q\|^2). \tag{19}$$

The first term is zero because both probability vectors normalize. Substituting $\Delta q = Jd + O(\|d\|^2)$ yields the result. □

Moment-information geometry. Let the rows of $A \in \mathbb{R}^{r \times m}$ assign r retained features to the m detector outcomes. Define the retained mean, its sensitivity, the single-draw covariance, and the resulting moment-information geometry by

$$\mu_A = Aq, \qquad J_A = AJ, \qquad \Sigma_A = A(D - qq^\top)A^\top, \qquad G_A = J_A^\top \Sigma_A^+ J_A. \tag{20}$$

This matrix is the covariance-whitened sensitivity of the empirical moment mean. It is the Godambe information of the optimal linear estimating function, and it is also the Fisher information of a local Gaussian moment experiment when Σ_A is held fixed. It is not, in general, the exact pushforward Fisher information of the full feature-valued observation $A(Z)$.

Proposition 3 (Tangent-space projection order). There is an orthogonal projector P_A such that, for $X = D^{-1/2} J$,

$$G_A = X^\top P_A X, \qquad F_{\mathrm{count}} = X^\top X. \tag{21}$$

Consequently, $0 \preceq G_A \preceq F_{\mathrm{count}}$. Equality holds exactly when $P_A X = X$, equivalently when every column of X lies in the covariance-supported row space of the centered features.

Proof. Set $u = D^{1/2}\mathbf{1} = \sqrt{q}$ and $C = I - uu^\top$. Since $u^\top X = \mathbf{1}^\top J = 0$, we have $CX = X$. With $B = AD^{1/2}$ and $M = BC$,

$$\Sigma_A = MM^\top, \qquad AJ = BX = BCX = MX. \tag{22}$$

Therefore

$$G_A = X^\top M^\top (MM^\top)^+ MX = X^\top P_A X, \tag{23}$$

where $P_A = M^\top (MM^\top)^+ M$ is the orthogonal projector onto the row space of M. Because $0 \preceq P_A \preceq I$, $X^\top P_A X \preceq X^\top X$. The equality condition follows from $X^\top (I - P_A) X = 0$. □

Proposition 4 (Full one-hot moment recovery). For $A = I_m$, the moment-information geometry equals the count Fisher: $G_A = F_{\mathrm{count}}$.

Proof. When $A = I_m$, $M = D^{1/2} C$. Left multiplication by the invertible matrix $D^{1/2}$ leaves its row space equal to $u^\perp$. The columns of X lie in $u^\perp$, so $P_A X = X$ and Proposition 3 gives the claim. □

This is a local projection statement about information retained by empirical moment means. A moment bank can remove a complete score direction, as in the rank-one example, without implying that the full feature-valued observation is globally insufficient for every task.

A.2 Same-count natural step and trust rescaling

For parameter j, let shifted detector-count arms provide empirical laws $\widehat{q}_{j,+}$ and $\widehat{q}_{j,-}$. In the studied two-parameter model, the exact shift identities are

$$\partial_{\theta_1} q = q_{1,+\pi/4} - q_{1,-\pi/4}, \qquad \partial_{\theta_2} q = \tfrac{1}{2}\left(q_{2,+\pi/2} - q_{2,-\pi/2}\right). \tag{24}$$

The base arm estimates q, and the four shifted arms estimate J. The task gradient and $\widehat{F}$ are deterministic functions of these same five multinomial count vectors.

EMA-aware dependent same-count bound. For the general d-parameter case, index the base and shifted arms by $a \in \mathcal{A}$, with $|\mathcal{A}| = A = 1 + 2d$. Arm a has categorical law $q_a \in \Delta^{m-1}$, empirical frequency f_a, and sample size n_a. For fixed $0 < \gamma < 1$, define

$$\widetilde{q}_a = (1-\gamma) f_a + \frac{\gamma}{m}\mathbf{1}, \qquad \bar{q}_a = (1-\gamma) q_a + \frac{\gamma}{m}\mathbf{1}. \tag{25}$$

The empirical and population-shrunken Jacobians have columns

$$\widehat{J}_{:,j} = r_j(\widetilde{q}_{j,+} - \widetilde{q}_{j,-}), \qquad \bar{J}_{:,j} = r_j(\bar{q}_{j,+} - \bar{q}_{j,-}). \tag{26}$$

Let

$$\widehat{F} = \widehat{J}^\top \operatorname{diag}(\widetilde{q}_0)^{-1} \widehat{J}, \quad \bar{F} = \bar{J}^\top \operatorname{diag}(\bar{q}_0)^{-1} \bar{J}. \tag{27}$$

For target categorical law τ, set $\widehat{g} = -\widehat{J}^\top(\tau/\widetilde{q}_0)$ and $\bar{g} = -\bar{J}^\top(\tau/\bar{q}_0)$, with componentwise division.

One-update concentration event. Let $A = 1 + 2d$, $n_{\min} = \min_a n_a$, $r_{\max} = \max_j |r_j|$, and define

$$e = (1-\gamma)\sqrt{\frac{\log(2Am/\delta)}{2n_{\min}}}, \qquad h = \frac{\gamma}{m}, \qquad E_J = 2r_{\max}\sqrt{md}\, e, \tag{28}$$

$$B_J = \|\bar{J}\|_2, \qquad E_F = \frac{(2B_J + E_J)E_J}{h} + \frac{B_J^2 e}{h^2}, \quad E_g = \frac{E_J}{h} + \frac{B_J e}{h^2}. \tag{29}$$

Hoeffding's inequality for every categorical indicator, followed by a union bound over the Am arm–outcome pairs, gives

$$\max_{a,z} |f_{a,z} - q_{a,z}| \le \sqrt{\frac{\log(2Am/\delta)}{2n_{\min}}} \tag{30}$$

with probability at least $1-\delta$. Multiplication by $1-\gamma$ gives $\max_a \|\widetilde{q}_a - \bar{q}_a\|_\infty \le e$. Both base laws have every component at least $h = \gamma/m$. Because each Jacobian column is a difference of two arms,

$$\|\widehat{J} - \bar{J}\|_{\mathrm{F}} \le 2r_{\max}\sqrt{md}\, e = E_J. \tag{31}$$

Write $\widehat{D} = \operatorname{diag}(\widetilde{q}_0)$ and $\bar{D} = \operatorname{diag}(\bar{q}_0)$. Adding and subtracting terms yields

$$\begin{aligned}\widehat{F} - \bar{F} = (\widehat{J} - \bar{J})^\top \widehat{D}^{-1}\widehat{J} + \bar{J}^\top \widehat{D}^{-1}(\widehat{J} - \bar{J}) \\ + \bar{J}^\top(\widehat{D}^{-1} - \bar{D}^{-1})\bar{J}.\end{aligned} \tag{32}$$

The support floor gives

$$\|\widehat{D}^{-1}\|_2 \le h^{-1}, \qquad \|\widehat{D}^{-1} - \bar{D}^{-1}\|_2 \le e h^{-2}. \tag{33}$$

Together with $\|\widehat{J}\|_2 \le B_J + E_J$, submultiplicativity gives

$$\|\widehat{F} - \bar{F}\|_2 \le \frac{(2B_J + E_J)E_J}{h} + \frac{B_J^2 e}{h^2} = E_F. \tag{34}$$

For the gradient, add and subtract $\bar{J}^\top(\tau/\widetilde{q}_0)$. Since $\|\tau\|_2 \le 1$, the same inverse-probability bounds imply

$$\|\widehat{g} - \bar{g}\|_2 \le E_J h^{-1} + B_J e h^{-2} = E_g. \tag{35}$$

Probability-smooth objectives. The categorical-KL calculation is a sharper instance of a general local bound. Let $g(q, J) = J^\top b(q)$, and suppose on the shrunken neighborhood that

$$\|b(q)\|_2 \le B_b, \qquad \|b(q) - b(q')\|_2 \le L_b \|q - q'\|_2. \tag{36}$$

Because $\|\widetilde{q}_0 - \bar{q}_0\|_2 \le \sqrt{m}\, e$,

$$\|\widehat{g} - \bar{g}\|_2 \le E_J B_b + B_J L_b \sqrt{m}\, e =: E_g^{\mathrm{gen}}. \tag{37}$$

Categorical KL has $b(q) = -\tau/q$ and admits the sharper Eq. (35). A weighted moment MSE has $b(q) = A^\top W(Aq - y)$; a fixed linear residual objective has the same form after composing A with its residual operator. Thus both are covered on any bounded local neighborhood. The Helmholtz instance uses a fixed discretized operator and feature map; this argument is a step-perturbation result, not a PDE-solver convergence theorem.

EMA expansion. At update s, apply the preceding conditional on the history before drawing that update's arms, with failure probability δ_s. Choose $\sum_{s=1}^{t} \delta_s \le \delta$. The union bound then gives one joint event on which all per-update metric and gradient bounds hold. Define

$$w_{t,1} = \prod_{r=2}^{t} \beta_r, \qquad w_{t,s} = (1 - \beta_s) \prod_{r=s+1}^{t} \beta_r, \quad 2 \le s \le t. \tag{38}$$

Unrolling the EMA recursion yields

$$\widehat{M}_t = \sum_{s=1}^{t} w_{t,s} \widehat{F}_s, \qquad \bar{M}_t = \sum_{s=1}^{t} w_{t,s} \bar{F}_s, \qquad \sum_{s=1}^{t} w_{t,s} = 1, \tag{39}$$

with the nonnegative weights in Eq. (38). Therefore

$$\|\widehat{M}_t - \bar{M}_t\|_2 \le \sum_{s=1}^{t} w_{t,s} \|\widehat{F}_s - \bar{F}_s\|_2 \le E_{M,t}. \tag{40}$$

This covers the first-observation initialization, constant or variable β_s, and $\beta_s = 0$. For multiple inputs with objective weights $\pi_x \ge 0$ and $\sum_x \pi_x = 1$, the same proof unions over input–arm–outcome triples and averages the input-wise radii with π_x.

Proof of Proposition 1. Equations (34)–(40) establish the joint EMA event. The resolvent identity gives

$$\begin{aligned} \widehat{v}_t - \bar{v}_t = & -(\widehat{M}_t + \lambda I)^{-1}(\widehat{g}_t - \bar{g}_t) \\ & + (\widehat{M}_t + \lambda I)^{-1}(\widehat{M}_t - \bar{M}_t)(\bar{M}_t + \lambda I)^{-1} \bar{g}_t. \end{aligned} \tag{41}$$

Both EMA matrices are positive semidefinite, so each damped inverse has operator norm at most $1/\lambda$. Apply the time-t gradient bound and Eq. (40). No factorization of the shared gradient and metric estimators is used. □

Proposition 5 (Detector-KL transfer for a general applied displacement). Let F_t be the current unshrunken detector Fisher, let $\bar{F}_t$ be its population-shrunken counterpart, and let $\widehat{M}_t^{\mathrm{full}}$ and $\bar{M}_t^{\mathrm{full}}$ be the empirical and population full-matrix EMAs. Let $\mathcal{S}_t$ be either the identity or diagonal extraction, and set $\widehat{T}_t = \mathcal{S}_t(\widehat{M}_t^{\mathrm{full}})$. Define

$$L_{\mathrm{EMA},t} = \|\bar{M}_t^{\mathrm{full}} - \bar{F}_t\|_2, \qquad B_{\mathrm{geom},t} = \|\mathcal{S}_t(\bar{F}_t) - \bar{F}_t\|_2, \qquad B_{\gamma,t} = \|\bar{F}_t - F_t\|_2, \tag{42}$$

and $\Gamma_t = E_{M,t} + L_{\mathrm{EMA},t} + B_{\mathrm{geom},t} + B_{\gamma,t}$. Suppose $\|\widehat{M}_t^{\mathrm{full}} - \bar{M}_t^{\mathrm{full}}\|_2 \le E_{M,t}$ and, for every $\|w\|_2 \le R$,

$$\left|\mathrm{KL}(q_{\theta_t} \,\|\, q_{\theta_t + w}) - \tfrac{1}{2} w^\top F_t w\right| \le \frac{L_{\mathrm{KL},t}}{6} \|w\|_2^3 \tag{43}$$

Then every such displacement satisfies Eq. (14). Consequently, the pre-box trusted proposal $\widetilde{u}_t$ in Eq. (11) satisfies Eq. (15), whereas the applied box-projected displacement a_t satisfies Eq. (14) with $\widehat{\delta}_t(a_t)$ in place of Δ_t.

Proof. Identity and diagonal extraction are nonexpansive in operator norm. The triangle inequality therefore gives

$$\begin{aligned}\|\widehat{T}_t - F_t\|_2 &\le \|\mathcal{S}_t(\widehat{M}_t^{\mathrm{full}}) - \mathcal{S}_t(\bar{M}_t^{\mathrm{full}})\|_2 \\ &\quad + \|\mathcal{S}_t(\bar{M}_t^{\mathrm{full}}) - \mathcal{S}_t(\bar{F}_t)\|_2 + B_{\mathrm{geom},t} + B_{\gamma,t} \\ &\le E_{M,t} + L_{\mathrm{EMA},t} + B_{\mathrm{geom},t} + B_{\gamma,t}. \end{aligned} \tag{44}$$

Therefore

$$\tfrac{1}{2} w^\top F_t w \le \widehat{\delta}_t(w) + \tfrac{1}{2}\Gamma_t \|w\|_2^2. \tag{45}$$

Adding the stated remainder proves Eq. (14); applying the trusted-proposal quadratic and norm cap proves Eq. (15). □

For full trust, $\mathcal{S}_t$ is the identity and $B_{\mathrm{geom},t} = 0$. For diagonal trust, $\mathcal{S}_t$ extracts the diagonal, so the same $E_{M,t}$ and $L_{\mathrm{EMA},t}$ bounds apply by nonexpansiveness. The calibration records $B_{\mathrm{geom},t}$ as the current population-Fisher full-to-diagonal gap, exactly as defined above.

Effect of the parameter box. Let $e_t = a_t - \widetilde{u}_t$ be the correction introduced by Eq. (12). Since $\theta_t \in \mathcal{C}$ and $\Pi_{\mathcal{C}}$ is Euclidean projection, $\|a_t\|_2 \le \|\widetilde{u}_t\|_2$. Positive semidefiniteness and the triangle inequality for the induced seminorm give

$$\widehat{\delta}_t(a_t) \le \left(\sqrt{\widehat{\delta}_t(\widetilde{u}_t)} + \sqrt{\widehat{\delta}_t(e_t)}\right)^2 \le \left(\sqrt{\Delta_t} + \sqrt{\widehat{\delta}_t(e_t)}\right)^2. \tag{46}$$

For diagonal $\widehat{T}_t$, coordinatewise projection contracts every weighted coordinate square, so $\widehat{\delta}_t(a_t) \le \widehat{\delta}_t(\widetilde{u}_t)$. A full positive-semidefinite quadratic need not contract: off-diagonal cancellation can disappear when only some coordinates are clipped. The full-metric applied step therefore uses the general transfer inequality rather than the Δ_t corollary.

Population lag and the variance–lag tradeoff. Equation (39) gives the deterministic history bound

$$L_{\mathrm{EMA},t} \le \sum_{s=1}^{t} w_{t,s} \|\bar{F}_s - \bar{F}_t\|_2. \tag{47}$$

If the population-shrunken Fisher is L_F-Lipschitz along the parameter path, then

$$L_{\mathrm{EMA},t} \le L_F \sum_{s=1}^{t} w_{t,s} \|\theta_s - \theta_t\|_2. \tag{48}$$

Larger decay spreads the weights over more historical batches: for centered, conditionally uncorrelated perturbations, the corresponding second moment contracts with $\sum_s w_{t,s}^2$, whereas Eqs. (47)–(48) generally grow with path history. The proposition uses the conservative weighted radii and does not require this variance model.

Same-count versus disjoint streams. With B samples spread equally over A arms, reuse provides $n_{\min} = \lfloor B/A \rfloor$ samples to both constructions. A disjoint split using fraction s for the gradient has at most $\lfloor sB/A \rfloor$ gradient and $\lfloor (1-s)B/A \rfloor$ metric samples per arm. For every $0 < s < 1$, both are no larger than the reuse count. Hence the two concentration radii entering the proposition are weakly smaller under reuse. This ordering concerns finite-sample perturbation; expected optimization loss additionally depends on the model, trajectory, and regularization.

Proposition 6 (Zero-additional-measurement trust projection). Let $d = -\alpha(\widehat{M} + \lambda I)^{-1}\widehat{g}$ with $\widehat{M} \succeq 0$, $\alpha > 0$, and $\lambda > 0$, and define $\widehat{\delta} = \frac{1}{2} d^\top \widehat{M} d$. For $\Delta > 0$, set

$$s = \begin{cases} 1, & \widehat{\delta} \le \Delta, \\ \sqrt{\Delta/\widehat{\delta}}, & \widehat{\delta} > \Delta, \end{cases} \qquad d_{\mathrm{trust}} = sd. \tag{49}$$

Then $\frac{1}{2} d_{\mathrm{trust}}^\top \widehat{M} d_{\mathrm{trust}} \le \Delta$, with equality when clipping occurs. If $\widehat{g}$ and the Fisher estimates entering $\widehat{M}$ are formed from the shared arms above, computing s requires no additional circuit or detector sample.

Proof. Positive semidefiniteness gives $\widehat{\delta} \geq 0$. If clipping occurs, homogeneity of the quadratic form gives $\frac{1}{2}(sd)^\top \widehat{M}(sd) = s^2\widehat{\delta} = \Delta$. Otherwise the original quadratic is already feasible. All quantities in s are algebraic functions of the estimated gradient, metric, and chosen radius. □

If the trusted step is subsequently capped at Euclidean radius $R > 0$, the cap multiplies it by a factor in $[0, 1]$ and therefore preserves the same quadratic bound.

The iteration-indexed schedule changes only Δ:

$$\Delta_t = \Delta_0 \left[1 - (1-\eta)\left(\frac{t/T - \rho}{1-\rho}\right)_+\right], \quad \rho = 0.75, \quad \eta = 0.10, \tag{50}$$

where the positive-part argument is clipped to $[0, 1]$. It therefore depends on neither exact loss nor validation measurements.

A.3 Optical and residual-noise objects

The minimal ideal state is

$$|\psi(\theta)\rangle = F_3 R_{12}(0.37) P_1(\theta_2) R_{01}(\theta_1)|100\rangle. \tag{51}$$

If $\pi_j = |\langle j|\psi\rangle|^2$ and detector efficiencies are $\zeta = (0.88, 0.61, 0.74)$, the retained no-click and click probabilities are

$$q_0 = \sum_{j=0}^{2}(1-\zeta_j)\pi_j, \qquad q_{j+1} = \zeta_j \pi_j. \tag{52}$$

Gaussian phase jitter replaces π by its expectation under an additive $\mathcal{N}(0, \sigma_\phi^2)$ phase; the implementation evaluates this expectation with fixed five-node Gauss–Hermite quadrature. Dark counts apply the replacement channel $q^{(d)} = (1-d)q + d(0, 1/3, 1/3, 1/3)^\top$.

For a retained scalar detector feature with independent input-wise variances v_x, let $\Sigma_u = \mathrm{diag}(v_x)$ and let R be the discretized strong-form Helmholtz operator including its boundary rows. The propagated residual covariance is

$$C_r = R\Sigma_u R^\top \succeq 0. \tag{53}$$

If J_r is the residual Jacobian, the residual-weighted metric $J_r^\top (C_r + \varepsilon I)^+ J_r$ is also positive semidefinite. These facts explain why high-frequency residuals amplify count-moment noise; they do not constitute a convergence theorem for a general PDE solver.

B Experimental, statistical, and reproducibility details

B.1 Observation models and prespecified design

All optimization experiments are CPU simulations with independent streams for target construction and finite-count sampling. Every method receives the exact target law as part of the synthetic fitting task. All count-geometry variants and Adam otherwise use sampled current-model detector laws online; the declared exact-QFIM comparator also receives an uncharged exact state metric. Exact current-model detector probabilities, losses, and population matrices provide offline evaluation and calibration. The development selector scores exact endpoint loss; its selected identifiers are fixed before holdout execution.

For a categorical detector law with m outcomes, arm counts c_a, and $N_a = \sum_z c_{a,z}$, every online estimator uses

$$\widetilde{q}_a = (1-\gamma)c_a/N_a + \gamma \mathbf{1}/m. \tag{54}$$

The fixed fraction γ makes the support floor independent of arm allocation. The high-dimensional and two-photon studies use $\gamma = 0.08$; the earlier equal-arm controls use the fraction equivalent to a Jeffreys pseudocount $a_0 = 1/2$ at their reference arm size. The accounting distinguishes emitted shots, detected and accepted events, no-click events, circuit evaluations, gradient-attributed shots, metric-attributed shots, shared shots, and offline exact oracle calls. A shared event is charged once to physical measurements even when it enters both estimators.

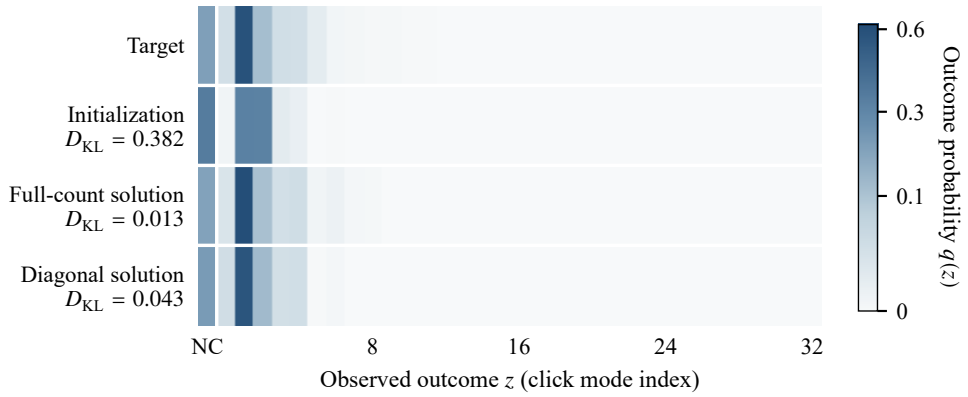


Figure 4: Observed single-photon solutions for one fixed, paired holdout case at $d = 32$. Columns are no-click (NC) followed by click modes 0–32; color encodes each exact detector-outcome probability $q(z)$. Rows show the target, shared initialization, full-count final solution, and diagonal final solution. The row annotations report $D_{\mathrm{KL}}(q_\star \| q)$. This individual case illustrates the shape of the solutions and is not a cohort-average result.

Single-photon graph construction. The d-parameter model begins in mode zero of $d+1$ real modes and applies d Givens rotations. A graph supplies the ordered rotation pairs. Chains couple adjacent modes; stars share mode zero; broom graphs interpolate a prescribed hub fraction; recursive trees are generated once from recorded graph seeds. Each rotation has the exact shift pair $(\pi/4, 1)$. The observation alphabet is no-click followed by a click in each mode.

Detector efficiencies have mean 0.72 and half-range 0.24. Alternating-low, alternating-high, blocked, and seeded-scrambled assignments use the same efficiency multiset and are independent of graph construction. At $d = 32$, one base and 64 shifted arms receive 64 emitted shots for each of 96 updates. A run therefore contains 399,360 emitted shots and 6,240 circuit evaluations.

Figure 4 shows what the fitted observation law looks like in one fixed holdout case. It includes the no-click outcome, so the displayed probabilities retain the detector loss that defines the observed geometry. The case is the lexicographically first target/design stream for the broom graph with a 0.10 hub fraction and the alternating-low detector profile; it was selected by seed, not by its final loss.

Prespecified panels and physical accounting. Table 2 records the complete new experiment suite. Across the seven panels it contains 18,496 runs, 7,163,825,152 emitted shots, and 110,994,016 circuit evaluations. The panel designs and comparison families were fixed before their results were analyzed.

Mechanism. FF, FD, DF, and DD use learning rate 0.40, damping 0.05, shrinkage 0.08, EMA decay 0.90, initial trust radius 0.003, terminal radius fraction 0.10, and Euclidean cap 0.35. The trusted proposal is subsequently projected into $[0.08, 1.45]^{32}$. Four graphs, four profiles, four targets, and eight design streams form 512 paired cells. The first letter always denotes the direction matrix and the second the trust matrix.

Equal method-wise tuning. FF and DD receive the same twelve-point balanced grid spanning learning rates $\{0.25, 0.40, 0.60\}$, damping $\{0.03, 0.07\}$, radii $\{0.0015, 0.0040\}$, EMA $\{0.75, 0.90\}$, and terminal fractions $\{0.05, 0.20\}$. Exact-QFIM NG receives the 4×3 product of learning rates $\{0.15, 0.30, 0.45, 0.60\}$ and damping $\{0.02, 0.05, 0.10\}$. Adam receives twelve learning rates from 0.003 to 0.24. Each candidate is evaluated on the same 96

Table 2: Prespecified experiment accounting. Emitted shots and circuit evaluations are totals over the panel.

Panel	Runs	Emitted shots	Circuit evaluations
$d = 32$ FF/FD/DF/DD mechanism	2,048	817,889,280	12,779,520
Matched-cost tuning development	4,608	1,840,250,880	28,753,920
Graph holdout after method-wise selection	9,216	3,680,501,760	57,507,840
Finite-count/EMA calibration	384	202,309,632	1,806,336
Reuse versus disjoint splits	576	179,850,240	3,224,160
Two-photon tuning development	512	136,314,880	2,129,920
Two-photon holdout	1,152	306,708,480	4,792,320
Total	18,496	7,163,825,152	110,994,016

development cells. Candidate scores first average log loss over design streams inside a graph/profile/target cell and then weight targets, profiles, and graphs equally. Candidates within 1% of the minimum enter a stability tie set; target-level log-loss standard deviation and candidate identifier break ties. The selected FF and DD schedules use $(\alpha, \lambda, \beta, \Delta_0, \eta) = (0.40, 0.03, 0.75, 0.004, 0.20)$ and $(0.25, 0.07, 0.75, 0.004, 0.05)$, respectively. Exact-QFIM natural gradient uses $(\alpha, \lambda) = (0.45, 0.10)$, and Adam uses learning rate 0.027.

The graph holdout contains broom hub fractions $\{0.10, 0.22, 0.37, 0.58, 0.73, 0.91\}$ and recursive graph seeds $\{15511, 15527, 15541, 15551, 15559, 15569\}$. Its target, design, graph, and scrambled-detector seeds do not occur in development. The selected schedules were fixed before holdout evaluation.

B.2 Statistical inference and finite-count checks

For graph k, profile p, target t, design stream s, and methods a, b, let $R_{kpts} = \log\{L^a_{kpts}/L^b_{kpts}\}$ with a fixed loss floor 10^{-15}. Point estimates exponentiate the equally weighted mean of R. Each of 50,000 bootstrap draws independently resamples targets and paired design streams while retaining every profile and graph. Contrasts within a stated family use the same draw. If $\widehat{r}_j^{(b)}$ is the log-ratio estimate in draw b, the simultaneous upper limit uses the 0.95 quantile of

$$\max_j \{\widehat{r}_j^{(b)} - \widehat{r}_j\}. \tag{55}$$

Familywise two-sided intervals analogously use the maximum absolute centered deviation. The analysis rejects duplicated keys, incomplete pairs, nonfinite losses, and mismatched shot or circuit-evaluation counts rather than filtering them. The $d = 32$ mechanism, tuned graph holdout, reuse, two-photon pooled contrasts, and graph-wise intervals form separate multiplicity families.

Table 3: Complete $d = 32$ factorial contrasts. The first letter denotes direction geometry and the second trust geometry.

Contrast	Ratio	95% CI	U_{95}	U_{95}^{sim}
Direction: FF / DF	0.757	[0.645, 0.871]	0.851	0.857
Direction: FD / DD	0.824	[0.758, 0.894]	0.881	0.933
Trust: FF / FD	0.969	[0.898, 1.041]	1.029	1.097
Trust: DF / DD	1.055	[0.955, 1.150]	1.133	1.195
Interaction	0.919	[0.813, 1.035]	1.010	1.040

Count reuse at matched physical cost. For a split fraction s, each observed arm first draws its usual multinomial count vector. A sequential multivariate-hypergeometric partition assigns exactly $\lfloor sN \rceil$ observations to the gradient subset and the remainder to the Fisher subset. The two partitions are disjoint and sum to the original arm count. Reuse supplies the original count vector to both estimators. This construction keeps circuit arms, emit-

Table 4: Fixed unseen-graph intervals for FF/DD after method-wise development selection. B1–B6 are broom graphs, with hub fractions shown in parentheses; T1–T6 are recursive trees, with graph seeds shown in parentheses. Familywise intervals use synchronized max statistics across all twelve graphs. Support is the fraction of paired cells in which the first method has lower loss.

Graph	Ratio	95% CI	Familywise 95% CI	Support
B1 (10% hub)	0.737	[0.619, 0.869]	[0.555, 0.978]	68.8%
B2 (22% hub)	0.705	[0.537, 0.939]	[0.531, 0.936]	71.4%
B3 (37% hub)	0.766	[0.677, 0.867]	[0.577, 1.017]	65.6%
B4 (58% hub)	0.737	[0.649, 0.855]	[0.555, 0.979]	70.8%
B5 (73% hub)	0.724	[0.609, 0.848]	[0.545, 0.961]	69.8%
B6 (91% hub)	0.722	[0.611, 0.845]	[0.544, 0.959]	69.3%
T1 (15511)	0.756	[0.697, 0.817]	[0.569, 1.004]	80.7%
T2 (15527)	0.714	[0.655, 0.770]	[0.537, 0.948]	85.4%
T3 (15541)	0.745	[0.676, 0.816]	[0.561, 0.989]	78.6%
T4 (15551)	0.868	[0.787, 0.949]	[0.654, 1.153]	62.0%
T5 (15559)	0.770	[0.691, 0.833]	[0.580, 1.022]	79.2%
T6 (15569)	0.691	[0.621, 0.766]	[0.521, 0.918]	80.2%

ted shots, update counts, and circuit evaluations fixed while changing only the information available to the two estimators.

The graph panel uses one recursive $d = 32$ graph, two detector profiles, four targets, and twelve design streams. Born and high-frequency fitting each use 24 paired seeds. Population gradient, Fisher, and natural-step errors are computed offline at the stored graph states. The complete endpoint family is reported below; it includes the low-dimensional contrasts whose signs differ across split fractions.

Table 5: Complete reuse/disjoint-split endpoint contrasts at matched physical cost. Each ratio compares reuse against the stated gradient:metric split. The multiplicity family contains all nine rows.

Panel	Contrast	Ratio	95% CI	U_{95}	U_{95}^{sim}
Born	25:75 split	0.396	[0.225, 0.685]	0.629	1.070
Born	50:50 split	2.043	[1.097, 3.869]	3.494	5.525
Born	75:25 split	0.687	[0.271, 1.906]	1.599	1.858
High frequency	25:75 split	0.769	[0.449, 1.309]	1.196	2.079
High frequency	50:50 split	1.306	[0.665, 2.765]	2.431	3.531
High frequency	75:25 split	0.149	[0.066, 0.368]	0.317	0.404
Graph	25:75 split	0.308	[0.271, 0.351]	0.343	0.834
Graph	50:50 split	0.456	[0.383, 0.541]	0.526	1.233
Graph	75:25 split	0.443	[0.370, 0.539]	0.522	1.199

Finite-count and EMA calibration. Calibration crosses $d \in \{16, 32\}$, full/diagonal geometry, two detector profiles, two targets, eight design streams, and 16/64/256 shots per arm. We analyze updates 1, 48, and 96. At each update the fixed analysis stores actual and theoretical errors for the Jacobian, current Fisher, population EMA, task gradient, and regularized natural step. Its KL fields evaluate the pre-box trusted proposal, even when the trajectory update is later projected into the parameter box. They store population EMA lag, trust-geometry approximation, shrinkage bias, the true Fisher quadratic, exact detector KL, the observed remainder, and an offline envelope. None of these population objects enters the optimizer.

The five estimator bounds attain empirical coverage one, but their magnitude is conservative. The KL envelope also covers every trusted proposal; because it uses exact population gaps and the realized exact remainder, it is an a-posteriori decomposition rather than an online certificate. The local Fisher quadratic is much sharper: the median exact-KL/quadratic

ratio is 0.999 and its 5–95% interval is 0.931–1.023. These calibration results apply to pre-box proposals; they do not assess the projected update.

Table 6: Finite-count calibration at 1,152 held-out optimization states. The first five rows compare estimator error with a concentration bound. The detector-KL row evaluates an offline envelope for the trusted proposal before box projection; it uses exact population terms and the realized remainder and is not an online certificate for the projected update.

Quantity	Median	90%	95%	99%	Maximum	Median envelope
Jacobian	0.465	1.067	1.171	1.286	1.382	16.746
Current Fisher	3.627	61.718	106.852	200.190	521.967	1.05×10^5
EMA Fisher	2.389	28.380	37.935	110.231	202.489	1.05×10^5
Task gradient	0.836	14.002	23.463	55.971	101.306	2.81×10^4
Natural step	1.141	5.563	7.727	13.127	22.589	1.38×10^7
True detector KL	2.25×10^{-4}	0.002	0.003	0.007	0.012	6.42×10^3

B.3 Two-photon model and generalized shift rule

Two photons enter modes zero and one of a six-mode real passive circuit. The ideal two-photon basis contains 21 occupations. Loss maps it to 28 observed occupations with zero, one, or two detected photons. To give the detector model common support, the lossy channel C_{loss} is composed with a fixed symmetric readout background,

$$C = (1-\varepsilon)C_{\text{loss}} + \frac{\varepsilon}{28}\mathbf{1}_{28}\mathbf{1}_{21}^{\top}, \qquad \varepsilon = 0.002. \tag{56}$$

This channel operation is part of the data-generating law; it is distinct from the online frequency shrinkage in Eq. (54).

For topology τ and detector profile ξ, the target is the exact detector law $r^{\tau,\xi} = q_{\theta^\star}^{\tau,\xi}$ and the objective is

$$\mathcal{L}^{\tau,\xi}(\theta;\theta^\star) = \mathrm{KL}\Big(q_{\theta^\star}^{\tau,\xi} \,\Big\|\, q_{\theta}^{\tau,\xi}\Big). \tag{57}$$

The exact target law enters the sampled KL gradient as the task label. Exact current-model detector losses provide offline trajectories and endpoints and score the development candidates; they do not influence online updates, returned iterates, or holdout model selection. For target key s_t, the eight components of $\theta^\star$ are drawn independently from $\mathrm{Unif}[0.25, 1.10]$ using key $s_t + 211$. For design key s_d, the paired key is $s = 10^6 s_d + s_t$. The initialization is $\theta_0 = \mathrm{clip}(\theta^\star + \delta, 0.10, 1.40)$, where the components of δ are drawn from $\mathrm{Unif}[-0.65, 0.65]$ and any magnitude below 0.24 is raised to 0.24 without changing its sign. Target, initialization, and count stream are paired across methods within every cell.

The ordered trainable rotation pairs are

$$\mathcal{R}_{\text{mesh}} = ((0,2),(1,3),(2,4),(3,5),(0,3),(1,4),(2,5),(0,5)), \tag{58}$$

$$\mathcal{R}_{\text{hub}} = ((0,2),(0,3),(1,2),(1,4),(0,5),(1,3),(2,4),(3,5)). \tag{59}$$

The three detector-efficiency vectors are

$$\eta_{\text{alternating}} = (0.58, 0.94, 0.58, 0.94, 0.58, 0.94), \tag{60}$$

$$\eta_{\text{blocked}} = (0.58, 0.58, 0.58, 0.94, 0.94, 0.94), \tag{61}$$

$$\eta_{\text{scrambled}} = (0.94, 0.94, 0.58, 0.94, 0.58, 0.58), \tag{62}$$

where the scrambled ordering is fixed by profile key 20611.

For a single trainable rotation, every outcome probability is a trigonometric polynomial of degree at most four. With nodes $\phi_r = 2\pi r/9$ and frequencies $k = -4, \ldots, 4$, the exact differentiation weights are

$$w_r = \frac{1}{9}\sum_{k=-4}^{4} ik\exp(-ik\phi_r), \qquad \partial_{\theta_j} q_\theta = \sum_{r=0}^{8} w_r q_{\theta+\phi_r e_j}. \tag{63}$$

The weights are real and $w_0 = 0$, so the separately sampled base arm plus eight nonzero shifts for each of eight parameters gives 65 arms. Analytical state derivatives and the nine-node rule agree to numerical precision in both mesh and hub circuits.

Development crosses two circuits, two profiles, two targets, four design streams, four methods, and four candidates per method. Table 7 reports the complete grid and marks the selected candidates. Here Δ_0 is the initial detector-KL radius, β is the metric-EMA decay, and η is its terminal fraction. All candidates use 64 shots per arm for 64 updates, frequency shrinkage 0.08, Euclidean cap $R = 0.35$, and the post-proposal box $[0.08, 1.45]^8$. FF and DD contract Δ_t from update fraction 0.75; exact-QFIM natural gradient and Adam have no detector-trust schedule. Adam uses $(\beta_1, \beta_2, \epsilon) = (0.9, 0.999, 10^{-8})$.

Table 7: Complete two-photon development grid. Bold identifiers mark the selected candidates; a dash denotes a quantity not used by that optimizer.

Method	ID	α	λ	β	Δ_0	η
FF/DD	f01	0.20	0.05	0.75	0.0015	0.10
FF/DD	f02	0.35	0.05	0.90	0.0030	0.10
FF/DD	f03	0.50	0.08	0.90	0.0015	0.20
FF/DD	f04	0.65	0.08	0.75	0.0030	0.20
Exact-QFIM NG	q01	0.15	0.03	–	–	–
Exact-QFIM NG	q02	0.30	0.05	–	–	–
Exact-QFIM NG	q03	0.45	0.08	–	–	–
Exact-QFIM NG	q04	0.60	0.10	–	–	–
Adam	a01	0.010	–	–	–	–
Adam	a02	0.025	–	–	–	–
Adam	a03	0.050	–	–	–	–
Adam	a04	0.100	–	–	–	–

The two-photon holdout retains the alternating and blocked development profiles, adds the scrambled profile, and introduces four new targets and twelve new design streams. It is simulation only and uses the same count histograms for the FF/DD gradient and metric estimates.

Table 8: Two-photon held-out simulation. The three pooled contrasts form one synchronized multiplicity family; only full count geometry versus exact-QFIM natural gradient has $U_{95}^{\mathrm{sim}} < 1$.

Contrast	Ratio [95% CI]	Familywise 95% CI	U_{95}^{sim}
Full / diagonal count	0.895 [0.772, 0.998]	[0.654, 1.227]	1.157
Full / exact-QFIM NG	0.653 [0.471, 0.882]	[0.476, 0.894]	0.843
Full / Adam	1.278 [1.057, 1.521]	[0.933, 1.751]	1.651

B.4 Solution view, earlier controls, and reproducibility

The two-photon outcome alphabet also permits a direct view of the recovered solution. Figure 5 uses the first target/design stream in the mesh/alternating holdout cell, again fixed by seed before looking at its loss.

The two-parameter controls cover Born-distribution fitting, high-frequency moment fitting, and low/high-frequency Helmholtz residuals. Born fitting uses 20,000 emitted shots, step size 0.08, and EMA 0.90. High-frequency fitting uses 25 inputs, 256,000 shots, step size 0.50, and no temporal aggregation. The Helmholtz controls use operator $\partial_{xx} + k^2$ and retain boundary rows; their purpose is to expose amplification of detector-moment noise, not to claim a general PDE solver.

The component suite contains exactly 80 configurations and 1,392 runs: 18 geometry configurations, 30 shot-scaling configurations, 11 trust schedules, and 21 detector-intervention configurations. Tables 9–12 list every configuration, including strong detector perturbations and configurations that do not improve the endpoint.

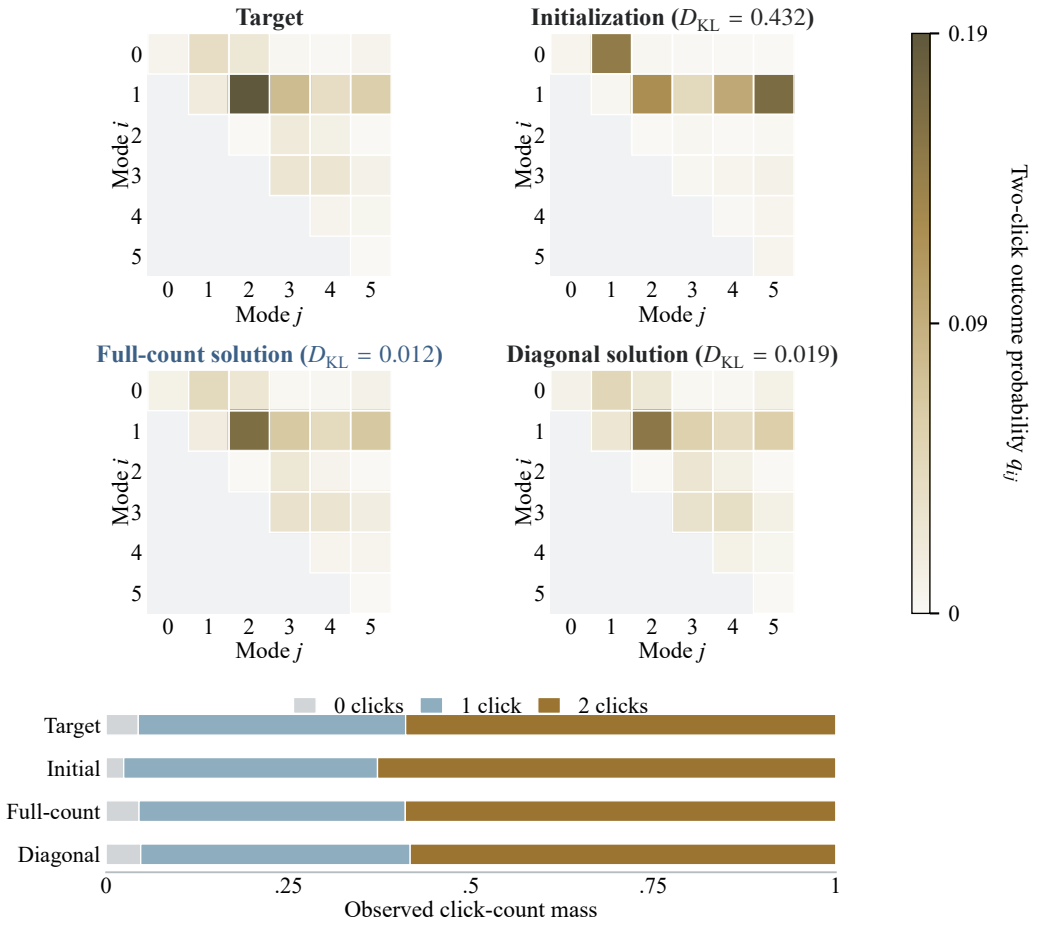


Figure 5: Observed two-photon solutions for one fixed, paired holdout case. The four upper-triangular heatmaps give the unconditioned probabilities q_{ij} of observing two clicks in modes $i \leq j$ (including two photons in one mode when $i = j$). The stacked bars retain the complementary no-click and one-click probability masses. The displayed KL values use the full 28-outcome law, rather than the heatmaps alone. This case illustrates individual solution structure and is not a cohort-average result.

Table 9: Complete component matrix: geometry. All configurations are reported.

Panel	Condition	Method	Median loss	Runs	Shots/run	Circuits/run
Geometry	Born	FockTrust	0.003	24	20,000	625
Geometry	Born	No EMA	0.004	24	20,000	625
Geometry	Born	No shrinkage	0.136	24	20,000	625
Geometry	Born	Diagonal	0.003	24	20,000	625
Geometry	Born	Isotropic	0.008	24	20,000	625
Geometry	Born	Moment	0.005	24	20,000	625
Geometry	Born	Exact count	0.001	24	20,000	625
Geometry	Born	Exact QFIM	0.003	24	20,000	625
Geometry	Born	Separate 4x	0.049	24	19,840	620
Geometry	High frequency	FockTrust	5.45×10^{-5}	24	256,000	8,000
Geometry	High frequency	No EMA	5.45×10^{-5}	24	256,000	8,000
Geometry	High frequency	No shrinkage	0.027	24	256,000	8,000
Geometry	High frequency	Diagonal	5.15×10^{-5}	24	256,000	8,000

Panel	Condition	Method	Median loss	Runs	Shots/run	Circuits/run
Geometry	High frequency	Isotropic	1.05×10^{-4}	24	256,000	8,000
Geometry	High frequency	Moment	7.55×10^{-5}	24	256,000	8,000
Geometry	High frequency	Exact count	5.55×10^{-5}	24	256,000	8,000
Geometry	High frequency	Exact QFIM	7.17×10^{-4}	24	256,000	8,000
Geometry	High frequency	Separate 4x	0.005	24	256,000	8,000

Table 10: Complete component matrix: shot scaling. All configurations are reported.

Panel	Condition	Method	Median loss	Runs	Shots/run	Circuits/run
Shots	Born, 8/arm	FockTrust	0.005	16	20,000	2,500
Shots	Born, 8/arm	Plug-in	0.015	16	20,000	2,500
Shots	Born, 8/arm	Exact QFIM	0.006	16	20,000	2,500
Shots	Born, 16/arm	FockTrust	0.004	16	20,000	1,250
Shots	Born, 16/arm	Plug-in	0.005	16	20,000	1,250
Shots	Born, 16/arm	Exact QFIM	0.006	16	20,000	1,250
Shots	Born, 32/arm	FockTrust	0.002	16	20,000	625
Shots	Born, 32/arm	Plug-in	0.009	16	20,000	625
Shots	Born, 32/arm	Exact QFIM	0.003	16	20,000	625
Shots	Born, 64/arm	FockTrust	0.003	16	19,840	310
Shots	Born, 64/arm	Plug-in	0.002	16	19,840	310
Shots	Born, 64/arm	Exact QFIM	0.002	16	19,840	310
Shots	Born, 128/arm	FockTrust	0.015	16	19,840	155
Shots	Born, 128/arm	Plug-in	0.011	16	19,840	155
Shots	Born, 128/arm	Exact QFIM	0.005	16	19,840	155
Shots	High frequency, 8/arm	FockTrust	1.96×10^{-4}	16	256,000	32,000
Shots	High frequency, 8/arm	Plug-in	1.96×10^{-4}	16	256,000	32,000
Shots	High frequency, 8/arm	Exact QFIM	1.97×10^{-4}	16	256,000	32,000
Shots	High frequency, 16/arm	FockTrust	1.09×10^{-4}	16	256,000	16,000
Shots	High frequency, 16/arm	Plug-in	1.09×10^{-4}	16	256,000	16,000
Shots	High frequency, 16/arm	Exact QFIM	2.33×10^{-4}	16	256,000	16,000
Shots	High frequency, 32/arm	FockTrust	7.46×10^{-5}	16	256,000	8,000
Shots	High frequency, 32/arm	Plug-in	7.46×10^{-5}	16	256,000	8,000
Shots	High frequency, 32/arm	Exact QFIM	0.001	16	256,000	8,000
Shots	High frequency, 64/arm	FockTrust	6.65×10^{-4}	16	256,000	4,000
Shots	High frequency, 64/arm	Plug-in	6.65×10^{-4}	16	256,000	4,000
Shots	High frequency, 64/arm	Exact QFIM	0.008	16	256,000	4,000
Shots	High frequency, 128/arm	FockTrust	0.007	16	256,000	2,000
Shots	High frequency, 128/arm	Plug-in	0.007	16	256,000	2,000
Shots	High frequency, 128/arm	Exact QFIM	0.043	16	256,000	2,000

Table 11: Complete component matrix: trust schedules. All configurations are reported.

Panel	Condition	Method	Median loss	Runs	Shots/run	Circuits/run
Trust	High frequency	No trust	73.317	24	128,000	4,000
Trust	High frequency	Constant	6.367	24	128,000	4,000
Trust	High frequency	Late 0.50	5.171	24	128,000	4,000
Trust	High frequency	Late 0.25	3.309	24	128,000	4,000
Trust	High frequency	Late 0.10	2.308	24	128,000	4,000
Trust	High frequency	Late 0.03	3.259	24	128,000	4,000
Trust	High frequency	Early 0.10	4.233	24	128,000	4,000
Trust	High frequency	Very late 0.10	4.927	24	128,000	4,000
Trust	Low frequency	No trust	5.459	12	63,360	1,980
Trust	Low frequency	Constant	1.958	12	63,360	1,980
Trust	Low frequency	Late 0.10	0.752	12	63,360	1,980

Table 12: Complete component matrix: detector interventions. All configurations are reported.

Panel	Condition	Method	Median loss	Runs	Shots/run	Circuits/run
Detector	Nominal	FockTrust	2.40×10^{-5}	12	256,000	8,000
Detector	Nominal	Exact count	4.51×10^{-5}	12	256,000	8,000
Detector	Nominal	Exact QFIM	0.001	12	256,000	8,000
Detector	Efficiency 0.8	FockTrust	4.50×10^{-5}	12	256,000	8,000
Detector	Efficiency 0.8	Exact count	1.27×10^{-5}	12	256,000	8,000
Detector	Efficiency 0.8	Exact QFIM	0.003	12	256,000	8,000
Detector	Efficiency 0.6	FockTrust	2.14×10^{-4}	12	256,000	8,000
Detector	Efficiency 0.6	Exact count	3.35×10^{-5}	12	256,000	8,000
Detector	Efficiency 0.6	Exact QFIM	0.008	12	256,000	8,000
Detector	Dark 0.01	FockTrust	4.69×10^{-5}	12	256,000	8,000
Detector	Dark 0.01	Exact count	5.74×10^{-5}	12	256,000	8,000
Detector	Dark 0.01	Exact QFIM	0.001	12	256,000	8,000
Detector	Dark 0.03	FockTrust	4.96×10^{-5}	12	256,000	8,000
Detector	Dark 0.03	Exact count	6.45×10^{-5}	12	256,000	8,000
Detector	Dark 0.03	Exact QFIM	0.002	12	256,000	8,000
Detector	Phase 0.05	FockTrust	2.50×10^{-5}	12	256,000	8,000
Detector	Phase 0.05	Exact count	6.53×10^{-5}	12	256,000	8,000
Detector	Phase 0.05	Exact QFIM	0.002	12	256,000	8,000
Detector	Phase 0.10	FockTrust	2.52×10^{-5}	12	256,000	8,000
Detector	Phase 0.10	Exact count	6.23×10^{-5}	12	256,000	8,000
Detector	Phase 0.10	Exact QFIM	0.002	12	256,000	8,000

Reproducibility files. The anonymous release contains run records, settings, selected schedules, and figure data sufficient to reconstruct the reported comparisons and physical measurement counts. Development and holdout targets and design streams are disjoint. Automated checks cover probability normalization, analytic versus shifted derivatives, positive-semidefinite geometry, moment projection, EMA weights, trust feasibility, shared/split accounting, and two-photon common support.