Jacobian-Velocity Bounds for Deployment Risk Under Covariate Drift

Jonathan R. Landers
jonathan.robert.landers@gmail.com
ORCID: 0000-0003-1872-6179

***Abstract*—We study long-horizon deployment of a frozen predictor under dynamic covariate shift. A time-domain Poincare inequality reduces temporal risk volatility to derivative energy, and a Jacobian-velocity theorem bounds that energy by the accumulated tangent amplification of the feature stream along the deployment path. When drift is low-rank, the leading term is governed by directional Jacobian energy in the drift subspace, motivating drift-aligned tangent regularization (DTR). Rather than smoothing the network isotropically, DTR penalizes sensitivity only along estimated drift directions, and the same geometric quantity yields a matched monitoring score. We validate the theorem-to-method pipeline in three experiments, namely a synthetic benchmark for the time-domain inequality, a controlled synthetic comparison against isotropic Jacobian regularization, and a field-deployment study on the UCI Air Quality sensor dataset. DTR consistently reduces risk volatility and directional gain relative to standard training and isotropic smoothing. Moderate drift-subspace misspecification is tolerable while orthogonal misspecification largely removes the benefit.**

I. Introduction

Distribution shift is often discussed statically by training on one distribution, testing on another, and measuring the gap [1], [2]. Long-horizon deployment is different. A model may be frozen while the environment moves continuously, so the relevant object is not a single out-of-domain error but a risk trajectory. Temporal benchmarks and deployment studies make the same point from different angles. Even when labels and architecture are fixed, performance can move materially over time [3]–[6]. The practical question is therefore not only whether shift occurs, but how violently performance can move once deployment begins.

This paper develops a geometric answer. If the feature stream follows a path X_t and the predictor has local tangent map $J_f(X_t)$, then instability is governed not by shift magnitude alone but by the directional interaction between the data velocity $\dot{X}_t$ and the model tangent geometry. The dangerous quantity is $J_f(X_t)\dot{X}_t$. The intuition is simple. The same deployment path can be almost harmless for one model and damaging for another, because only motion through steep tangent directions accumulates into large score variation. If the stream moves through a flat direction, drift can be present without being operationally dangerous. If it moves through a steep direction, even gentle but persistent motion can compound into long-horizon degradation. The paper's central bound formalizes this. Under mild regularity and a domination condition on the performance field,

$$\mathrm{Var}_U(r(U)) \le \frac{\beta^2 T}{\pi^2} \int_0^T \mathbb{E}\|J_f(X_t)\dot{X}_t\|^2 dt.$$

First, a time-domain Poincare inequality converts temporal volatility into derivative energy. Second, a Jacobian-velocity theorem bounds that energy under explicit regularity assumptions on the deployment performance field. Third, a low-rank drift specialization makes a drift-aligned regularizer and a matched monitoring score almost inevitable. Fourth, we test the resulting story with a synthetic theorem sanity check, a compact isotropic-versus-directional comparison, and a real field-deployment study on Air Quality. The empirical sequence mirrors the logical one. It first verifies the inequality itself, then compares competing regularizers in the idealized low-rank regime, and only then moves to a real deployment path where the drift subspace must be estimated. Concretely, the paper contributes a defensible bound on risk volatility for frozen deployment, a low-rank reduction that isolates the drift subspace, a drift-aligned tangent regularizer (DTR), a matched hazard score, and empirical evidence that the same directional geometry matters. DTR reduces risk volatility and directional gain relative to standard training and isotropic smoothing in both controlled synthetic experiments and a real field-deployment study.

The paper is narrow by design. It does not address general distribution shift or test-time adaptation, but asks specifically what local geometry governs frozen-model degradation and whether that geometry remains visible on real drifting data.

II. Related Work

The paper sits closest to temporal distribution shift and dynamic deployment evaluation. Classical dataset-shift taxonomies separate covariate, label, and concept shift [1], [2], while concept-drift work emphasizes sequential adaptation under evolving streams [3]. More recent benchmark-driven work has made temporal structure explicit, both in naturally shifted datasets and in stylized train-deploy protocols [4]–[7]. Our focus is narrower. Rather than estimating a worst-case shift gap, we study the volatility of one frozen model along a deployment path.

The paper is also adjacent to test-time adaptation [8]–[10]. Those methods update the model online to track changing data. Here the predictor remains frozen. That restriction is deliberate. It isolates the geometric mechanism of degradation itself and yields a quantity that can be analyzed, regularized,

and monitored without mixing in the dynamics of online parameter updates.

Finally, the work draws on two technical lines. The regularization side is related to Jacobian-based control of sensitivity and robustness [11]–[13], but the object here is directional rather than isotropic. Only tangent energy in likely drift directions enters the bound. The monitoring side is related to shift detection, uncertainty under shift, and sequential change detection [14]–[20]. Our hazard score is not presented as a sequentially optimal detector. It is a model-aware proxy matched to the same drift geometry used in training.

III. Setup and Assumptions

Let $t \in [0, T]$ index deployment time. Let $X_t \in \mathbb{R}^d$ denote the covariate process, and let $f_\theta : \mathbb{R}^d \to \mathbb{R}$ be a frozen scalar score function. We study a scalar deployment performance field $g_\theta : \mathbb{R}^d \to \mathbb{R}$ and the associated trajectory

$$r(t) := \mathbb{E}[g_\theta(X_t)].$$

For classification, a natural example is

$$g_\theta(x) = \mathbb{E}_{Y \sim P_0(\cdot|x)}[\ell(f_\theta(x), Y)],$$

but the theorem below does not rely on this representation alone. Instead, we state the regularity needed to make the risk-trajectory argument defensible. The point of this setup is to separate two roles that are often conflated in shift analyses. The model is frozen, while the covariate stream is allowed to move.

To measure temporal instability, draw $U \sim \mathrm{Unif}[0, T]$ and define

$$\mathrm{Var}_U(r(U)) = \mathbb{E}[(r(U) - \mathbb{E}r(U))^2].$$

Assumption A1 (trajectory regularity). The sample paths $t \mapsto X_t$ are almost surely absolutely continuous and

$$\mathbb{E} \int_0^T \|\dot{X}_t\|^2 dt < \infty.$$

Assumption A2 (chain rule under the expectation). The field g_θ is weakly differentiable, $t \mapsto g_\theta(X_t)$ is almost surely absolutely continuous, and

$$g_\theta(X_t) - g_\theta(X_s) = \int_s^t \nabla g_\theta(X_u)^\top \dot{X}_u \, du$$

for all $0 \le s < t \le T$ on almost every sample path, with

$$\mathbb{E} \int_0^T |\nabla g_\theta(X_t)^\top \dot{X}_t| dt < \infty.$$

Assumption A3 (directional domination by the score Jacobian). There exists $\beta > 0$ such that for almost every x and every $v \in \mathbb{R}^d$,

$$|\nabla g_\theta(x)^\top v| \le \beta \|J_f(x) v\|.$$

Assumption A3 is the paper's main compatibility condition. It should not be read as a consequence of dynamic covariate shift alone. Rather, it asserts that along deployment directions, the local change in the chosen performance field is dominated by the local change in the score.

Remark 1 (When A3 holds and when it fails)**.** *A3 holds in the natural composition case. If $g_\theta(x) = h(f_\theta(x))$, then $\nabla g_\theta(x) = h'(f_\theta(x)) \nabla f_\theta(x)$ and A3 follows with $\beta = \sup |h'|$. For Bernoulli cross-entropy against a fixed soft target, $|h'| \le 1$, giving $\beta = 1$.*

A3 does not follow from pure covariate shift alone. The general risk $g_\theta(x) = \mathbb{E}_{Y \sim P_0(\cdot|x)}[\ell(f_\theta(x), Y)]$ carries an extra $\nabla \eta(x)$ term from the label conditional, which need not be dominated by $\|J_f(x)v\|$ even when η is time-invariant. A3 also fails under simultaneous concept shift, and at score saturation wherever $\nabla f_\theta(x) = 0$ but $\nabla g_\theta(x) \neq 0$.

IV. Main Theorem Package

The first step is purely temporal.

Lemma 1 (Time-domain derivative-energy control)**.** *If r is absolutely continuous on $[0, T]$, then*

$$\mathrm{Var}_U(r(U)) \le \frac{T}{\pi^2} \int_0^T (r'(t))^2 dt.$$

Proof. Let $\bar{r} = T^{-1} \int_0^T r(t) \, dt$. The Wirtinger inequality yields

$$\int_0^T (r(t) - \bar{r})^2 dt \le \frac{T^2}{\pi^2} \int_0^T (r'(t))^2 dt.$$

Divide by T. □

Lemma 1 says volatility is impossible without derivative energy. The next result identifies the geometric driver of that energy.

Theorem 1 (Jacobian-velocity control of risk volatility)**.** *Assume A1–A3 and suppose f_θ is a ReLU network, so $J_f(x)$ exists almost everywhere. Then r is absolutely continuous and*

$$\mathrm{Var}_U(r(U)) \le \frac{\beta^2 T}{\pi^2} \int_0^T \mathbb{E}\|J_f(X_t)\dot{X}_t\|^2 dt.$$

Proof. Fix $0 \le s < t \le T$. By Assumption A2 and Fubini,

$$\begin{aligned} r(t) - r(s) &= \mathbb{E}\left[\int_s^t \nabla g_\theta(X_u)^\top \dot{X}_u \, du\right] \\ &= \int_s^t \mathbb{E}[\nabla g_\theta(X_u)^\top \dot{X}_u] \, du. \end{aligned}$$

Hence r is absolutely continuous and, for almost every t,

$$r'(t) = \mathbb{E}[\nabla g_\theta(X_t)^\top \dot{X}_t].$$

Jensen's inequality and Assumption A3 then give

$$\begin{aligned} (r'(t))^2 &\le \mathbb{E}[(\nabla g_\theta(X_t)^\top \dot{X}_t)^2] \\ &\le \beta^2 \mathbb{E}\|J_f(X_t)\dot{X}_t\|^2. \end{aligned}$$

Apply Lemma 1. □

Theorem 1 isolates a single instability mechanism. Deployment volatility is driven by accumulated tangent amplification along the realized motion of the data stream. Figure 1 is a geometric rendering of that statement. In that sense, the theorem turns Jacobian-based sensitivity control [11]–[13] into a statement about temporal deployment risk rather than static

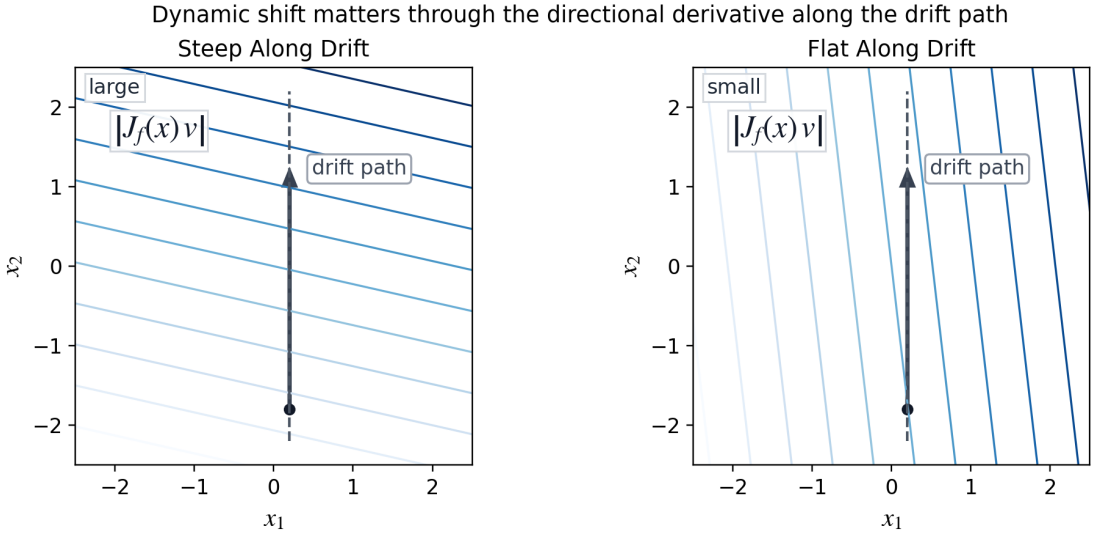

Fig. 1. Geometric intuition for drift-aligned instability. The deployment path is the same in both panels, but the local tangent geometry is different. When the drift direction aligns with a steep tangent direction, the score changes rapidly, whereas when the model is flat along that same direction, the score remains stable. Figure 1 visualizes the directional quantity that appears in Theorem 1.

local robustness. The bound also has a useful operational reading. Volatility requires two ingredients at once, data motion and tangent gain. If either one is small, long-horizon fluctuation remains small.

The low-rank regime gives the algorithmic specialization.

Corollary 1 (Drift-subspace bound)**.** *Assume the hypotheses of Theorem 1. If*

$$\dot{X}_t = Va_t + \rho_t, \qquad V \in \mathbb{R}^{d\times k},$$

with $V^\top V = I_k$ *and* $V^\top \rho_t = 0$*, then*

$$\mathrm{Var}_U(r(U)) \le \frac{2\beta^2 T}{\pi^2}(B_V + B_\rho),$$

where

$$B_V := \int_0^T \mathbb{E}[\|J_f(X_t)V\|_F^2 \|a_t\|^2]dt,$$
$$B_\rho := \int_0^T \mathbb{E}\|J_f(X_t)\rho_t\|^2 dt.$$

Proof. By Theorem 1,

$$\|J_f(X_t)\dot{X}_t\|^2 = \|J_f(X_t)Va_t + J_f(X_t)\rho_t\|^2$$
$$\le 2\|J_f(X_t)Va_t\|^2 + 2\|J_f(X_t)\rho_t\|^2.$$

Since V has orthonormal columns,

$$\|J_f(X_t)Va_t\|^2 \le \|J_f(X_t)V\|_F^2\|a_t\|^2.$$

Integrate over t. □

Corollary 1 is the paper's main reduction. When B_ρ is small, the leading stability term is governed by the directional Jacobian energy in the drift subspace rather than by isotropic smoothness. This is precisely the regime in which low-dimensional temporal drift and low-rank change models are empirically plausible [5], [18].

V. Method

The theorem package suggests a simple design principle. If instability is created by motion through steep directions, then regularization should flatten the model only where deployment is expected to move, not everywhere at once.

A. Drift-aligned tangent regularization

Let V estimate the dominant drift subspace from unlabeled deployment covariates. Two minimal estimators are sufficient for the present paper. One is a mean-difference direction

$$\Delta\mu_t = \mu_t - \mu_{t-\Delta}, \qquad v_t = \frac{\Delta\mu_t}{\|\Delta\mu_t\|},$$

or the top-k principal directions of the difference cloud $\{X_t - X_{t-\Delta}\}$ over a rolling window.

With a fixed subspace estimate V, we train using

$$\mathcal{L}_{\mathrm{DTR}}(\theta) = \mathbb{E}_{(X,Y)}\ell(f_\theta(X), Y) \;+\; \lambda\mathbb{E}_X\|J_f(X)V\|_F^2.$$

This is not global Jacobian smoothing. It is a directional motion constraint. The model is flattened only along directions expected to dominate future drift. Corollary 1 explains why this is the quantity worth shrinking, and it distinguishes DTR from isotropic Jacobian penalties that aim at broader smoothness or margin control [11]–[13]. In other words, DTR tries to buy smoothness exactly where the future trajectory is expected to travel, rather than spending regularization budget uniformly over directions the deployment path may never visit.

B. Monitoring score

The same geometry yields a matched deployment diagnostic. Let

$$s_t := \|\Delta\mu_t\|/\Delta, \qquad g_t := \mathbb{E}\|J_f(X_t)V_t\|_F^2, \qquad h_t := s_t^2 g_t.$$

When residual drift energy is small and V_t tracks the dominant motion, h_t is a model-aware hazard proxy. It becomes large only when the stream is moving quickly *and* the predictor is steep along that motion. Relative to generic shift alarms [14], [15] and sequential detection procedures [16], [17], [20], the point of h_t is alignment rather than optimality. It monitors exactly the directional mechanism singled out by the theory. This symmetry between training and monitoring is deliberate. The same geometry used to suppress future instability is reused to measure when future instability is likely.

VI. Experiments

A. Synthetic theorem sanity check

The synthetic study is kept compact but fully specified. It uses the smallest setup that exposes the paper's mechanism. There is one stable signal coordinate, one drifting nuisance coordinate, and a frozen classifier whose tangent geometry can either amplify or suppress that drift. This kind of temporally evolving evaluation is consistent with recent temporal-shift benchmarks, but here the drift direction is deliberately known so the theorem-to-method link can be read cleanly [5], [6]. Labels are sampled as $Y \sim \mathrm{Bernoulli}(1/2)$ and we write $S = 2Y - 1 \in \{-1, +1\}$. Conditional on S, features are generated as

$$x_1 \sim \mathcal{N}(1.05S, 0.90^2), \qquad x_2 \sim \mathcal{N}(1.25S + \delta(t), 0.55^2).$$

The first coordinate is a stable signal. The second is a nuisance coordinate that is predictive at training time but drifts during deployment. The label mechanism is held fixed throughout. Only $\delta(t)$ changes. This benchmark is intentionally instrumental rather than realistic. Its job is to make the paper's directional mechanism observable with as little confounding structure as possible.

The drift schedule is monotone and smooth, with speed

$$\dot{\delta}(t) = 0.55 + 1.05e^{-((t-0.33)/0.10)^2} + 0.75e^{-((t-0.76)/0.08)^2},$$
$$\delta(0) = 0,$$

so the stream contains two periods of rapid motion. We train a one-hidden-layer ReLU classifier with 12 hidden units, binary cross-entropy loss, Adam with learning rate 0.02, and 140 epochs. Each training set has 1024 examples drawn at $t = 0$, each deployment time uses a fresh sample of 2048 points, and we sweep $\lambda \in \{0, 0.01, 0.03, 0.08\}$ over two random seeds. DTR uses the true drift direction $V = e_2$ so that this experiment isolates the theorem-to-method link rather than subspace-estimation error.

Figure 2 plays the theorem-linked role. Each point is one trained model, plotted by empirical risk volatility versus the derivative-energy quantity from Lemma 1. The points sit below the diagonal, as they should, and DTR moves the cloud toward the lower-left corner. Across all eight models, mean volatility falls from 3.13×10^{-3} to 2.62×10^{-4}, while mean directional gain falls from 44.9 to 1.91.

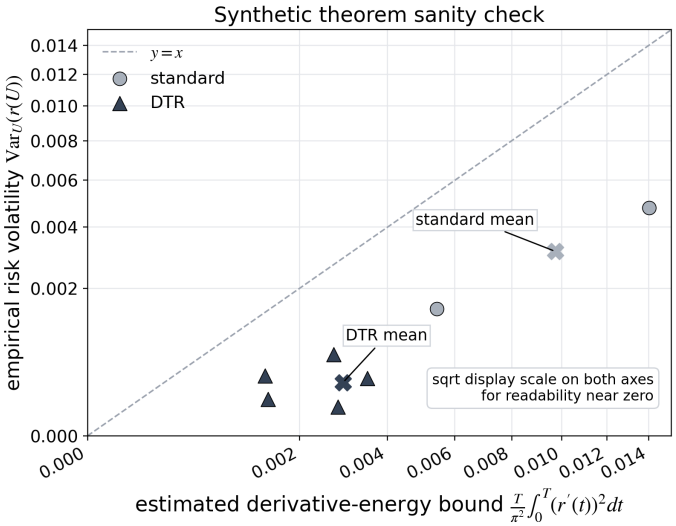

Fig. 2. Synthetic theorem sanity check. Each point is one trained model evaluated over the drifting deployment path. The horizontal axis is the derivative-energy quantity from Lemma 1, and the vertical axis is empirical volatility $\mathrm{Var}_U(r(U))$. DTR shifts models toward lower energy and lower volatility.

TABLE I
Directional comparison showing the mean over the nonzero-λ sweep. Lower is better on all metrics.

Method	Deriv. energy	Volatility	Dir. gain	Term. risk
Standard	4.80×10^{-3}	3.13×10^{-3}	44.9	0.171
Isotropic	6.11×10^{-4}	4.46×10^{-4}	2.15	0.168
DTR	3.14×10^{-4}	2.62×10^{-4}	1.91	0.134

B. Directional versus isotropic smoothing

The theorem suggests a sharper empirical question than whether "Jacobian regularization helps." When drift is concentrated in one known direction, does directional smoothing beat isotropic smoothing? This is the paper's main algorithmic comparison, because it asks whether the low-rank reduction is merely descriptive or whether it really changes the design choice. We answer that question in the same synthetic environment by replacing the DTR penalty with an isotropic Jacobian penalty

$$\lambda \mathbb{E}\|\nabla f_\theta(X)\|^2$$

while keeping the directional penalty

$$\lambda \mathbb{E}\|J_f(X)V\|_F^2$$

for DTR, with the same architecture, optimizer, seeds, and sweep $\lambda \in \{0.01, 0.03, 0.08\}$. Under rank-1 drift, isotropic smoothing spends budget in directions the deployment path does not substantially traverse, whereas DTR spends that budget only along the realized motion. Table I summarizes the mean values across the nonzero-λ sweep. DTR improves on both baselines on every metric. Figure 3 (left) shows the matched $\lambda = 0.03$ comparison.

We then test drift-subspace misspecification by rotating the one-dimensional subspace used by DTR. Writing the estimated direction as

$$\hat{v}_\alpha = (\sin\alpha, \cos\alpha),$$

its alignment with the true drift direction e_2 is $\cos\alpha$. With the correct subspace, the directional advantage remains strongest. With a mild 20° rotation, volatility rises by a factor of 1.43 and terminal risk by a factor of 1.20 relative to aligned DTR, but both remain far below the standard model. With an orthogonal subspace, the effect largely disappears. Derivative energy rises by a factor of 31.7, volatility by 23.9, and directional gain by 19.4. Figure 3 (right) therefore sharpens the paper's claim in the intended narrow way. DTR is more targeted than isotropic Jacobian smoothing when the drift geometry is right, and its gains depend on keeping the estimated drift subspace reasonably aligned with realized motion.

C. Air Quality field deployment

We next test whether the same directional story survives on a real field dataset with documented drift. The role of this experiment is different from the synthetic studies. It is not a clean theorem check or a controlled comparison, but a test of whether the same geometry still leaves a measurable trace once the drift subspace must be estimated from data. The UCI Air Quality dataset records hourly outputs of a five-sensor gas array together with temperature and humidity in an Italian city, and the associated field-study paper explicitly notes both concept drift and sensor drift [21]. We predict the reference CO concentration $\mathrm{CO(GT)}$ from the five sensor channels $\mathrm{PT08.S1}$–$\mathrm{PT08.S5}$ together with T, RH, and AH. Rows with the dataset's -200 missing-value marker are removed, leaving 1573 training points from the first 12 weeks, 580 validation points from the next 4 weeks, and 5191 deployment points from the remaining horizon, grouped into 20 biweekly deployment blocks.

To keep the protocol aligned with the theorem, the predictor is frozen after training and deployment risk is the blockwise mean-squared error trajectory. We estimate a fixed two-dimensional drift subspace V from unlabeled deployment covariates by taking the top singular vectors of consecutive biweekly mean-shift vectors. We then train a two-hidden-layer ReLU regressor with width 32, Adam with learning rate 0.01, and $\lambda \in \{0, 0.003, 0.01, 0.03, 0.08\}$ over three random seeds. Monitoring uses the same geometry as the method. If μ_t is the standardized covariate mean in block t, then

$$s_t = \|\mu_t - \mu_{t-1}\|, \qquad v_t = \frac{\mu_t - \mu_{t-1}}{\|\mu_t - \mu_{t-1}\|},$$
$$g_t = \mathbb{E}[(\nabla f_\theta(X_t)^\top v_t)^2], \qquad h_t = s_t^2 g_t.$$

The real-data result supports the same mechanism, even though the theorem is not claimed as an exact model of this field dataset. Across the full nonzero-λ sweep, mean deployment-risk volatility falls from 9.29×10^{-2} to 8.16×10^{-2}, while mean directional gain falls from 8.65×10^{-2} to 6.80×10^{-2}. The representative moderate setting $\lambda = 0.03$ is stronger. Across three seeds, mean volatility falls to 6.49×10^{-2} and mean directional gain to 6.09×10^{-2}, with lower mean terminal risk (0.149 versus 0.169). Figure 4 shows one representative seed. The standard regressor begins at blockwise MSE 0.363, peaks at 1.15, and ends at 0.176. The DTR regressor begins at 0.342, peaks at 0.949, and ends lower at 0.152. Its mean hazard score also falls from 0.212 to 0.148, indicating that the same directional quantity used in training remains informative at monitoring time.

VII. Discussion and Limitations

The theorem is mathematically narrow, by design. The crucial caution is Assumption A3. Dynamic covariate shift by itself does not imply that the conditional-risk field is controlled by the score Jacobian. We impose that domination explicitly because it is exactly what the theorem uses.

The three empirical pieces play different roles. The theorem sanity check asks whether the time-domain inequality is visible in trained models at all. The isotropic-versus-directional comparison asks whether the low-rank specialization has real design content. The Air Quality study asks whether the same directional quantities still matter once the setting becomes noisy and the drift subspace must be estimated. That progression is narrower than a benchmark paper, but it is also the right level of evidence for the claim being made here.

The low-rank story also depends on estimating a drift subspace. The synthetic misspecification study suggests that moderate angular error is tolerable but large misalignment is not. In the Air Quality study we estimate a fixed subspace from unlabeled future covariates, which is more realistic than the aligned synthetic setting but still retrospective. A real online deployment would require rolling estimation of V_t, and the quality of the monitoring score would then depend on residual drift energy and subspace-estimation error. Finally, the Jacobian-velocity bound is an upper bound. It is useful because it is actionable, not because it is tight in every regime, and it is best interpreted as a design guide for regularization and monitoring rather than as a complete description of all failure modes.

VIII. Conclusion

Long-horizon robustness under dynamic shift is best viewed as a tangent-control problem. The data stream contributes a velocity, the model contributes a local linear map, and deployment volatility is governed by their interaction. The central message is simple. Deployment becomes unstable when environmental motion repeatedly passes through directions to which the model is locally sensitive. The theorem package makes this precise, DTR acts on it, and the three experiments test whether the same directional geometry remains operative from clean synthetic settings through to noisy real sensor-drift data — it does, which is the paper's main empirical finding.

That positioning matters. Relative to temporal-shift benchmarks and static shift analyses [1], [4]–[6], the paper offers a concrete mechanism for why deployment trajectories differ across models exposed to the same evolving environment.

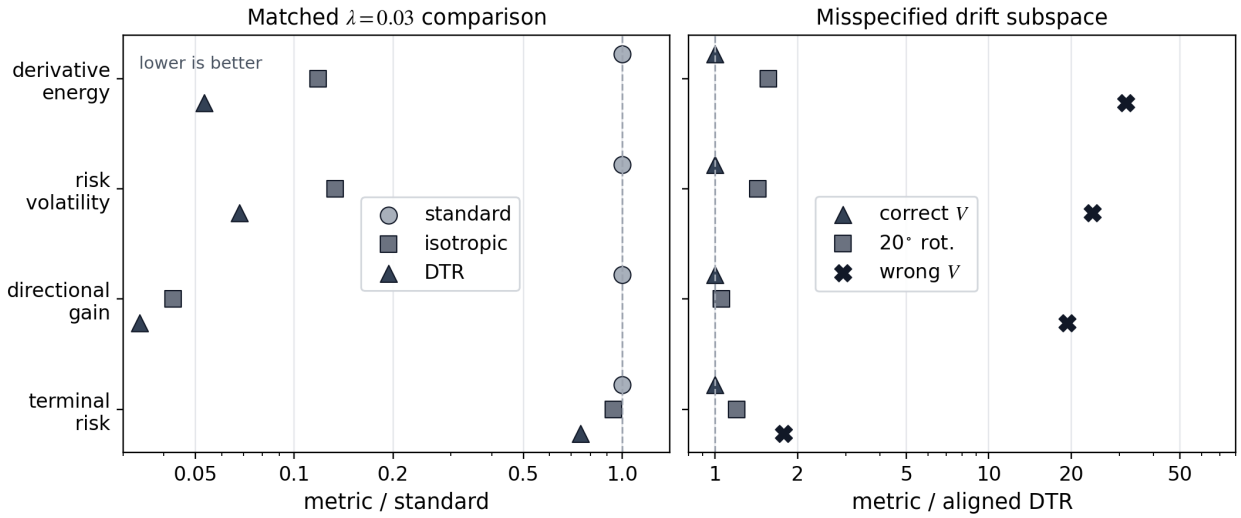

Fig. 3. Synthetic directional comparisons. Left shows matched $\lambda = 0.03$ ratios for derivative energy, risk volatility, directional gain, and terminal risk, normalized by the standard model. DTR is lower than isotropic Jacobian regularization on all four metrics. Right shows the same metrics normalized by aligned DTR for the correct subspace, a mild 20° rotation, and a wrong subspace. Mild misspecification is tolerable, but wrong subspaces are not.

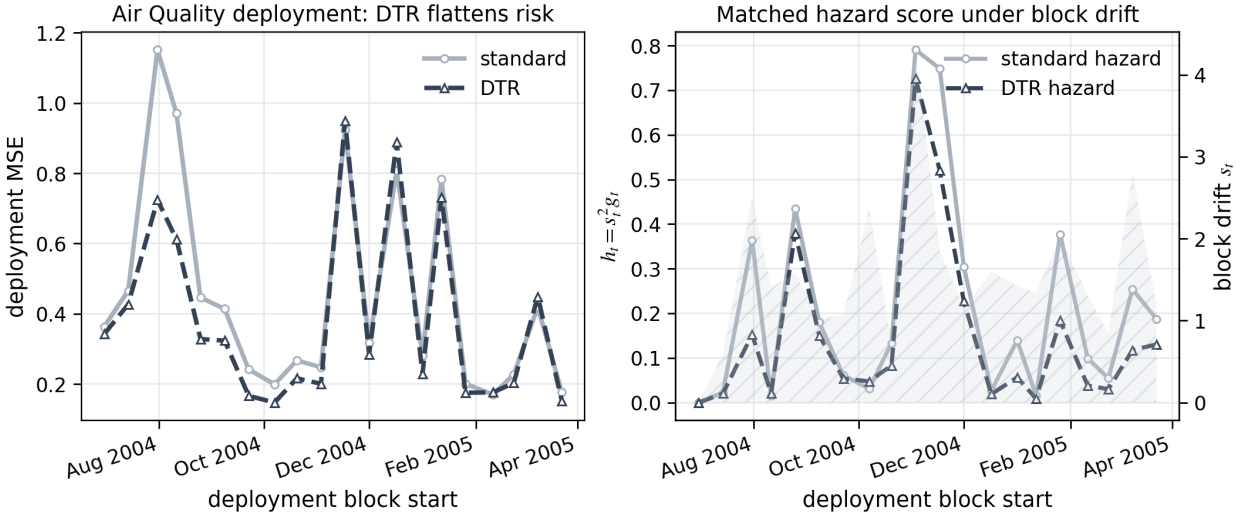

Fig. 4. Real field deployment on UCI Air Quality. Left shows biweekly deployment MSE for a standard CO regressor and a DTR regressor after training on the first 12 weeks and freezing the model. Right shows the matched hazard score $h_t = s_t^2 g_t$, with block drift magnitude in the background. DTR suppresses tangent gain along realized motion and therefore flattens both risk and hazard.

Relative to test-time adaptation [8]–[10], it studies the frozen-model regime in which geometric control is especially valuable. Relative to Jacobian smoothing and shift detection [11], [13], [14], [20], it ties training-time regularization and deployment-time monitoring to one directional notion of drift sensitivity. If future deployment motion is concentrated in a few directions, then robustness should be enforced, measured, and monitored in those directions rather than spread uniformly over the ambient space.

References

[1] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1–2, pp. 151–175, 2010.

[2] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodriguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[3] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 44:1–44:37, 2014.

[4] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, I. Haque, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang, "Wilds: A benchmark of in-the-wild distribution shifts," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 5637–5664.

[5] H. Yao, C. Choi, B. Cao, Y. Lee, P. W. Koh, and C. Finn, "Wild-time: A benchmark of in-the-wild distribution shift over time," in *Advances in Neural Information Processing Systems*, 2022.

[6] E. Han, C. Huang, and K. Wang, "Model assessment and selection under temporal distribution shift," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235, 2024, pp. 17 374–17 392.

[7] M. Simchowitz, A. Ajay, P. Agrawal, and A. Krishnamurthy, "Statistical learning under heterogeneous distribution shift," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, 2023, pp. 31 800–31 851.

[8] Y. Sun, X. Wang, Z. Liu, J. Miller, A. A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119, 2020, pp. 9229–9248.

[9] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *International Conference on Learning Representations*, 2021.

[10] S. Gui, X. Li, and S. Ji, "Active test-time adaptation: Theoretical analyses and an algorithm," in *International Conference on Learning Representations*, 2024.

[11] J. Sokolic, R. Giryes, G. Sapiro, and M. R. D. Rodrigues, "Robust large margin deep neural networks," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4265–4280, 2017.

[12] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein, "Sensitivity and generalization in neural networks: An empirical study," *International Conference on Learning Representations*, 2018.

[13] T. Kim and H. Yang, "An infinite-width analysis on the jacobian-regularised training of a neural network," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235, 2024, pp. 24 584–24 657.

[14] S. Rabanser, S. Günnemann, and Z. C. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," in *Advances in Neural Information Processing Systems*, 2019.

[15] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, 2019.

[16] S. Wu, E. Diao, J. Ding, T. Banerjee, and V. Tarokh, "Robust quickest change detection for unnormalized models," in *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, vol. 216, 2023, pp. 2314–2323.

[17] Y. Cao and Y. Xie, "Multi-sensor gradual change detection," in *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing*, 2015, pp. 827–834.

[18] Y. Xie and L. M. Seversky, "Sequential low-rank change detection," in *Proceedings of the 54th Annual Allerton Conference on Communication, Control, and Computing*, 2016, pp. 128–133.

[19] V. Krishnamurthy and L. Snow, "Quickest change detection using time inconsistent anticipatory and quantum decision modeling," in *Proceedings of the 58th Annual Allerton Conference on Communication, Control, and Computing*, 2022, pp. 1–8.

[20] A. Cooper and S. Meyn, "Quickest change detection using mismatched cusum," in *Proceedings of the 60th Annual Allerton Conference on Communication, Control, and Computing*, 2024, pp. 1–7.

[21] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sensors and Actuators B: Chemical*, vol. 129, no. 2, pp. 750–757, 2008.