POSTERIOR DRIFT PROFILES FOR ADAPTIVE GENERALIZATION AND CONCENTRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reusing data to choose and evaluate a model or query can introduce selection bias and alter sampling fluctuations. We introduce *posterior drift profiles*, which measure the largest Wasserstein displacement of a data law under log-likelihood perturbations of bounded sensitivity. A profile bound has two uses: its value supplies the selection-bias allowance in transferring sample accuracy to population accuracy, while its integral yields concentration around the fresh-data mean, including after stable selection. Both guarantees use the original law's profile, so the same calculation can serve different selection procedures and statistics. We prove exact tensorization for product laws under additive metrics, even when selection induces posterior dependence. Explicit calculations cover bounded queries, sampling without replacement, and unbounded Lipschitz queries under metric privacy. The nonlinear Gamma profile shows why the profile's shape matters: increasing likelihood sensitivity can enlarge both bias and fluctuation bounds. A one-sided extension also accommodates asymmetric, dataset-dependent sensitivity.

1 INTRODUCTION

A model's performance on observed data is useful when it predicts performance on new examples. For a predictor fixed independently of the evaluation sample, concentration inequalities control how far its empirical performance can deviate from its population performance. In practice, the same data often help choose the predictor: researchers compare models, tune hyperparameters, and revise analyses after inspecting validation scores. Reliable evaluation then requires accounting for both selection bias and the fluctuations that remain after selection.

Generalization theory controls the gap between sample and population performance through, for example, uniform convergence (Bartlett & Mendelson, 2002), algorithmic stability (Bousquet & Elisseeff, 2002), and information-theoretic bounds (Xu & Raginsky, 2017). Adaptive data analysis asks how to provide population-accurate answers when earlier answers guide later queries (Dwork et al., 2015; Bassily et al., 2016). A standard route combines answers already accurate on the sample with a generalization guarantee controlling the additional error from data reuse. We seek a common distributional bound that supports both this accuracy transfer and concentration inequalities before and after stable selection.

Selection changes the conditional data law. Let $X \sim P$ be the dataset and let the *transcript* Π be the observable record that determines the statistic f_Π: a returned predictor, or the released query–answer history and final selection. Conditional on $\Pi = \pi$, the statistic is fixed but the data follow $Q_\pi = \mathcal{L}(X \mid \Pi = \pi)$. The population benchmark is $\mu_\pi = \mathbb{E}_{X' \sim P} f_\pi(X')$, with the selected function held fixed and X' fresh. Selection bias is therefore the expectation shift $\mathbb{E}_{Q_\pi} f_\pi - \mathbb{E}_P f_\pi$.

For a dataset distance d and transcript log-likelihood $g(x) = \log p(\pi \mid x)$, Bayes' rule gives $Q_\pi = Q_g := e^g P / \mathbb{E}_P e^g$. We introduce the *posterior drift profile*

$$\Theta_{P,d}(\eta) = \sup_{\mathrm{Lip}_d(g) \le \eta} W_{1,d}(Q_g, P).$$

Here the supremum is over well-defined tilts with finite first moment. The condition $\mathrm{Lip}_d(g) \le \eta$ says that nearby datasets give similar likelihoods to the same output. Since Wasserstein distance

is the largest expectation shift of a 1-Lipschitz statistic, the profile bounds the drift of every L-Lipschitz statistic by $L\Theta_{P,d}(\eta)$. It separates a property of the sampling law and distance, (P, d), from the sensitivities of the selection procedure and statistic.

1.1 Main results and contributions

One profile, two uses. The profile value gives the selection-bias allowance in a transfer theorem: population error is at most conditional sample error plus $L\Theta_{P,d}(\eta)$ (Theorem 1). Integrating the same profile also gives

$$\log \mathbb{E}_{Q_g} e^{\lambda(f-\mathbb{E}_P f)} \le \int_0^\lambda L\Theta_{P,d}(\eta + uL)\, du \qquad (\lambda \ge 0),$$

whenever the integral is finite (Theorem 2). The derivative of the log-moment generating function is the expectation drift under the further tilt Q_{g+uf}. Setting $g = 0$ gives ordinary concentration; starting from a transcript likelihood gives concentration after selection. Only the original law's profile is needed, even if the posterior is dependent or no longer belongs to the original distribution family.

Exact tensorization and explicit examples. For product data and an additive metric $d_a = \sum_i a_i d_i$, we prove $\Theta_{P,d_a}(\eta) = \sum_i a_i \Theta_{P_i,d_i}(a_i\eta)$ (Theorem 3). This equality preserves nonlinear coordinate profiles and makes the method computable. We provide explicit bounds on several interesting examples. Bounded coordinates recover the familiar bounded-differences sampling term with an explicit selection allowance. Sampling without replacement demonstrates applicability to dependent data. Gaussian and Gamma profiles give guarantees for unbounded Lipschitz queries under metric privacy, where stability measures the magnitude of a numerical change. The Gamma example is especially informative: selection can enlarge both the bias and the fluctuation parameters, which a single global linear bound cannot capture (Section 5.4).

Beyond a fixed metric, and relation to existing methods. A one-sided local extension handles state-dependent coordinate weights and gives adaptive upper-tail concentration in Talagrand's local geometry (Section 6). Posterior arguments already underlie adaptive-data-analysis transfer (Jung et al., 2020; Shenfeld & Ligett, 2023) and Wasserstein generalization bounds (Rodríguez-Gálvez et al., 2021). Our contribution is the reusable likelihood-sensitivity profile, its exact tensorization, and the passage from expectation drift to concentration along further tilts. The advantage is a common calculation across sampling laws and geometries. We also prove that entropy and transportation inequalities can themselves bound the profile (Section 7). This means that, using our drift profile notion as a bridge, entropy and transportation methods can directly prove adaptive concentrations. On the other hand, with the nice properties developed here, bounding drift profile explicitly provides a new way of proving (adaptive) concentration inequalities. Appendix G gives detailed comparisons; Appendix H extends transfer to approximate DP on product data.

2 Posterior drift profiles

Let P have finite first moment on a Polish metric space (Ω, d). Write $\mathrm{Lip}_d(f) \le L$ when $|f(x) - f(y)| \le Ld(x, y)$. Lipschitz conditions refer to finite real-valued versions on the support of P. A score g defines $Q_g = e^g P/Z_g$, where $Z_g = \mathbb{E}_P e^g$; the tilt is *admissible* if $0 < Z_g < \infty$ and Q_g has finite first moment. Such moments ensure integrability of Lipschitz statistics.

Definition 1 (Posterior drift profile)**.** *The posterior drift profile of* (P, d) *is*

$$\Theta_{P,d}(\eta) := \sup_{\substack{g:\,\mathrm{Lip}_d(g)\le\eta \\ Q_g \text{ admissible}}} W_{1,d}(Q_g, P), \qquad \eta \ge 0. \tag{1}$$

The profile is nondecreasing, satisfies $\Theta_{P,d}(0) = 0$, and may be infinite. It depends on the original data law and metric, independently of a particular selection procedure.

Proposition 1 (Expectation drift)**.** *For an admissible* Q_g *with* $\mathrm{Lip}_d(g) \le \eta$ *and* $\mathrm{Lip}_d(f) \le L$*,*

$$|\mathbb{E}_{Q_g} f - \mathbb{E}_P f| \le L\Theta_{P,d}(\eta). \tag{2}$$

Proof. Kantorovich–Rubinstein duality gives $W_{1,d}(Q,P) = \sup_{\mathrm{Lip}_d(h)\le 1} |\mathbb{E}_Q h - \mathbb{E}_P h|$. Apply this to f/L when $L > 0$; constant statistics have zero drift. □

Throughout, $L\Theta(\eta)$ is defined to be zero when $L = 0$, even if $\Theta(\eta) = \infty$. A profile bound is a uniform drift allowance, not necessarily the actual shift of a particular statistic. Neither application below requires independence of P.

3 USE 1: TRANSFERRING SAMPLE ACCURACY TO THE POPULATION

Let $X \sim P$ as above and let Π be a standard Borel transcript determining k queries $f_{j,\Pi}$ and answers $a_j(\Pi)$. Assume the queries are jointly measurable in (x,π), and the answers and finite nonnegative sensitivity bounds are measurable in π. Put $\mu_j(\pi) = \mathbb{E}_{X'\sim P} f_{j,\pi}(X')$, with the selected function fixed. Conditional on the transcript, the answer and query are fixed while their sample evaluation remains random. Sample accuracy is an input to the theorem; drift controls the additional population error.

Suppose the mechanism has jointly measurable likelihood densities against a common σ-finite measure. On one set of transcripts of full marginal probability, assume finite log-likelihood versions satisfy

$$g_\pi(x) = \log p(\pi \mid x), \qquad \mathrm{Lip}_d(g_\pi) \le \eta(\pi). \tag{3}$$

Bayes' rule identifies the posterior, and finite first moments hold almost surely by the tower property:

$$Q_\pi(dx) = \mathcal{L}(X \mid \Pi = \pi)(dx) = \frac{p(\pi \mid x)}{\mathbb{E}_P p(\pi \mid X)} P(dx) = Q_{g_\pi}(dx). \tag{4}$$

Pure metric DP (Chatzikokolakis et al., 2013) requires $K_x(A) \le e^{\epsilon d(x,y)} K_y(A)$ for transcript laws K_x. It supplies $\eta = \epsilon$ when likelihood versions obey this bound simultaneously on the support of P outside one null set. Standard DP uses Hamming distance.

Theorem 1 (Posterior-drift transfer)**.** *In this transcript setting, let* $\mathrm{Lip}_d(f_{j,\pi}) \le L_j(\pi)$ *and set* $B_j(\pi) = L_j(\pi)\Theta_{P,d}(\eta(\pi))$. *For almost every* π*,*

$$|a_j(\pi) - \mu_j(\pi)| \le \mathbb{E}[|a_j(\pi) - f_{j,\pi}(X)| \mid \Pi = \pi] + B_j(\pi). \tag{5}$$

If the answers and queries lie in $[0,1]$ *and, for* $\alpha, \beta \in [0,1]$*,*

$$\mathbb{P}\left[\max_{j\le k} |a_j(\Pi) - f_{j,\Pi}(X)| > \alpha\right] \le \beta, \tag{6}$$

then, for every $\gamma \in (0,1)$*,*

$$\mathbb{P}_\Pi[\exists j :\ |a_j(\Pi) - \mu_j(\Pi)| > \alpha + \gamma + B_j(\Pi)] \le \beta/\gamma. \tag{7}$$

Proof. Insert the posterior expectation between the answer and its population benchmark:

$$|a_j - \mu_j| \le |a_j - \mathbb{E}_{Q_\pi} f_{j,\pi}| + |\mathbb{E}_{Q_\pi} f_{j,\pi} - \mu_j| \le \mathbb{E}_{Q_\pi}|a_j - f_{j,\pi}| + B_j.$$

This proves (5). Let $q(\pi)$ be the conditional probability of the joint sample-failure event in (6). Since $\mathbb{E}q(\Pi) \le \beta$, Markov's inequality gives $q(\Pi) \le \gamma$ except on a set of probability β/γ. On the remaining transcripts, every conditional sample error is at most $\alpha + (1-\alpha)q(\pi) \le \alpha + \gamma$. Substitute into (5). □

One conditional failure probability controls all answers, so there is no additional union bound over k; the assumed sample-accuracy guarantee may itself depend on k. This uses the posterior decomposition of Jung et al. (2020), with the sampling law and geometry entering through the profile. The expectation comparison also applies to unbounded Lipschitz queries. Corollary 2 below gives a tail-based transfer variant without the range restriction. Appendix H treats approximate DP through an expected excess drift.

4 USE 2: CONCENTRATION FROM POSTERIOR DRIFT

Starting from Q_g, multiplication by $e^{\lambda f}$ changes the score to $g + \lambda f$, whose sensitivity is at most $\eta + \lambda L$. Applying the original law's drift bound along this path controls fluctuations as well as bias. Define

$$\Psi_{\eta,L}(\lambda) = \int_0^\lambda L\Theta_{P,d}(\eta + uL)\, du, \qquad \lambda \geq 0, \qquad \Psi_{\eta,0} = 0. \tag{8}$$

Theorem 2 (Concentration by integrating posterior drift)**.** *Let Q_g be admissible with $\mathrm{Lip}_d(g) \leq \eta$, and let $\mathrm{Lip}_d(f) \leq L$. For every $\lambda \geq 0$ with $\Psi_{\eta,L}(\lambda) < \infty$,*

$$\log \mathbb{E}_{Q_g} e^{\lambda(f - \mathbb{E}_P f)} \leq \Psi_{\eta,L}(\lambda). \tag{9}$$

Consequently, for $r \in \mathbb{R}$,

$$Q_g[f - \mathbb{E}_P f \geq r] \leq \inf_{\lambda \geq 0:\, \Psi_{\eta,L}(\lambda) < \infty} e^{-\lambda r + \Psi_{\eta,L}(\lambda)}. \tag{10}$$

The same statements hold for $-f$.

Proof. First take bounded f. The density of Q_{g+uf} relative to Q_g is bounded above and below by positive constants, so these tilts remain admissible and differentiation under the integral is valid. Then

$$\frac{d}{du} \log \mathbb{E}_{Q_g} e^{u(f - \mathbb{E}_P f)} = \mathbb{E}_{Q_{g+uf}} f - \mathbb{E}_P f \leq L\Theta_{P,d}(\eta + uL).$$

Integrating from 0 to λ gives (9). For general Lipschitz f, apply this bound to $f_R = (-R) \vee (f \wedge R)$, which remains L-Lipschitz. Integrability gives $\mathbb{E}_P f_R \to \mathbb{E}_P f$; Fatou's lemma proves (9), including finiteness of the exponential moment. Chernoff's inequality gives (10). □

When $\Theta_{P,d}(\eta)$ and the integral are finite, write $B_{\eta,L} = L\Theta_{P,d}(\eta)$. Then

$$\Psi_{\eta,L}(\lambda) = \lambda B_{\eta,L} + \int_0^\lambda L\big[\Theta_{P,d}(\eta + uL) - \Theta_{P,d}(\eta)\big]\, du. \tag{11}$$

The initial value $B_{\eta,L}$ is a bias allowance; growth of the profile beyond η determines the remaining fluctuation bound. In particular, if $\Theta(u) \leq Cu$ for all $u \geq 0$, integrating this upper bound gives $\Psi_{\eta,L}(\lambda) \leq CL\eta\lambda + CL^2\lambda^2/2$. The quadratic term is unchanged by the starting sensitivity η. A nonlinear profile can instead record an increase in fluctuation scale, as in Section 5.4.

Corollary 1 (Concentration after stable selection)**.** *In the transcript setting of Section 3, let f_π be a jointly measurable selected statistic with $\mathrm{Lip}_d(f_\pi) \leq L_\pi$ and $\mu_\pi = \mathbb{E}_P f_\pi$. For almost every transcript π,*

$$\log \mathbb{E}[e^{\lambda(f_\pi(X) - \mu_\pi)} \mid \Pi = \pi] \leq \Psi_{\eta(\pi),L_\pi}(\lambda) \tag{12}$$

whenever the right-hand side is finite. If $\Theta(u) \leq Cu$ for all $u \geq 0$, with $C > 0$, then, for $L_\pi, t > 0$,

$$\mathbb{P}[|f_\pi(X) - \mu_\pi| \geq CL_\pi\eta(\pi) + t \mid \Pi = \pi] \leq 2e^{-t^2/(2CL_\pi^2)}. \tag{13}$$

Apply the theorem to the conditional law Q_{g_π} and optimize each one-sided Chernoff bound at $\lambda = t/(CL_\pi^2)$. Taking $g = 0$ gives *ordinary concentration*; conditioning gives concentration around the *fresh-data* mean without a functional inequality for the posterior.

Reading the bound as a generalization guarantee. For $0 < \delta < 1$, choose the conditional tail level in (13) and average over the transcript:

$$\mathbb{P}\Big[|f_\Pi(X) - \mu_\Pi| > CL_\Pi\eta(\Pi) + L_\Pi\sqrt{2C\log(2/\delta)}\Big] \leq \delta. \tag{14}$$

The two sensitivities may depend on the observed output; a constant statistic has zero deviation. For a selected predictor w_π and loss ℓ, suppose the empirical-risk function below satisfies $\mathrm{Lip}_d(f_\pi) \leq L_\pi$. Taking $f_\pi(X) = n^{-1}\sum_i \ell(w_\pi, X_i)$ on an i.i.d. sample makes $f_\pi(X)$ its empirical risk and $\mu_\pi = \mathbb{E}_Z \ell(w_\pi, Z)$ its population risk. Equation (14) then gives a confidence band around the

reused-sample evaluation, with an explicit selection allowance. Section 5.2 gives the bounded-loss specialization.

The same interpretation applies to model selection through feedback on an independent validation set. Condition on the training data and require the likelihood bound for the full validation transcript determining the chosen model, for almost every fixed training dataset. The number of candidates does not appear separately in the bound, but more interaction may increase the transcript sensitivity. The guarantee concerns the final selected model; simultaneous realized sample deviations of all candidate models require a further argument.

Combining concentration with accuracy transfer. A transcript whose answer is far from the population mean must either have a large sample error or force its selected query into a conditional tail. This connects the two uses of the profile and removes the bounded-range requirement from high-probability transfer, at the cost of controlling the profile along the full tilt path.

Corollary 2 (Transfer using conditional tails)**.** *In the transcript setting of Section 3, suppose real-valued answers and Lipschitz queries satisfy* (6) *for some* $\alpha \geq 0$ *and* $\beta \in [0,1]$*. Assume* $\Theta_{P,d}(u) \leq Cu$ *for all* $u \geq 0$ *with* $C > 0$*. For* $q \in (0,1)$*, define*

$$R_j(\pi) = CL_j(\pi)\eta(\pi) + L_j(\pi)\sqrt{2C\log(1/q)}.$$

Then

$$\mathbb{P}_\Pi[\exists j : |a_j(\Pi) - \mu_j(\Pi)| > \alpha + R_j(\Pi)] \leq \frac{\beta}{1-q}. \tag{15}$$

Proof. On a transcript violating the asserted accuracy threshold, choose the first violating index j and put $s = \operatorname{sign}(a_j - \mu_j)$. Both choices depend only on the transcript. If joint sample accuracy holds, then

$$s\big(f_{j,\pi}(X) - \mu_j(\pi)\big) \geq |a_j(\pi) - \mu_j(\pi)| - |a_j(\pi) - f_{j,\pi}(X)| > R_j(\pi).$$

The conditional one-sided bound gives probability at most q to this event; for a constant query it is impossible. Thus the conditional probability of sample failure is at least $1 - q$ on every violating transcript. Averaging and using (6) proves (15). □

The violating index is chosen from the transcript, so there is no additional union bound over queries. Taking $q = 1/2$ gives failure probability 2β. Unlike Theorem 1, this result uses a concentration radius but allows unbounded answers and queries. A nonlinear profile can be used in the same argument whenever its integrated bound supplies the required conditional tail radius.

5 Computing profiles: one estimate, two applications

The following calculations make the framework explicit. Tensorization reduces product laws to coordinates; the slice handles dependence; Gaussian and Gamma laws distinguish fixed from increasing fluctuation scales (Table 1). We prove tensorization below. For the remaining models we give proof ideas; complete proofs are in Appendices B–D and E.

Table 1: Profile calculations, with $m = \min(K, N-K)$ and Gamma rate β. The Hamming row assumes finite or countable coordinates; Section 5.2 treats general standard Borel spaces.

Reference law and metric	Profile $\Theta_{P,d}(\eta)$	What the example demonstrates
Product law, Hamming	$\leq n\tanh(\eta/4)$	Bounded-query benchmark
Uniform K-subset, swap	$\leq m\tanh(\eta/4)$	Dependent reference data
$N(0,\sigma^2 I)$, Euclidean	$= \sigma^2\eta$	A linear profile
$\mathrm{Gamma}(\alpha,\beta)$, Euclidean	$= \dfrac{\alpha\eta}{\beta(\beta-\eta)}, \eta < \beta$	Selection can increase the fluctuation scale

5.1 EXACT TENSORIZATION FOR PRODUCT REFERENCE LAWS

Let $P = \bigotimes_{i=1}^n P_i$ and $d_a(x,y) = \sum_i a_i d_i(x_i, y_i)$, with $a_i > 0$ and each P_i of finite first moment on a Polish metric space (Ω_i, d_i).

Theorem 3 (Product tensorization)**.** *For every $\eta \geq 0$,*

$$\Theta_{P,d_a}(\eta) = \sum_{i=1}^n a_i \Theta_{P_i,d_i}(a_i \eta). \tag{16}$$

More generally, if $Q \ll P$ has finite first moment and its conditional law $Q_i(\cdot \mid x_{<i})$ has a log-density relative to P_i with Lipschitz constant at most a deterministic $\rho_i \geq 0$ for almost every posterior prefix, then

$$W_{1,d_a}(Q,P) \leq \sum_i a_i \Theta_{P_i,d_i}(\rho_i). \tag{17}$$

Proof. For an admissible score g with $\mathrm{Lip}_{d_a}(g) \leq \eta$, integrate out the suffix under its prior law:

$$W_i(x_{<i}, u) = \int e^{g(x_{<i},u,z_{>i})}\, P_{>i}(dz_{>i}).$$

The conditional density of Q_g relative to P_i is proportional to W_i. The score bound gives $W_i(x_{<i},u) \leq e^{\eta a_i d_i(u,v)} W_i(x_{<i},v)$. Interchanging u, v shows that the conditional score is ηa_i-Lipschitz, for almost every posterior prefix.

To prove (17), take a bounded 1-Lipschitz test function h and its prior averages

$$H_i(x_{\leq i}) = \mathbb{E}_{P_{>i}} h(x_{\leq i}, X_{>i}), \qquad D_i = H_i - H_{i-1}, \qquad H_0 = \mathbb{E}_P h.$$

For each prefix, $D_i(x_{<i}, \cdot)$ is P_i-centered and a_i-Lipschitz. Conditional first moments are finite almost surely by disintegration, so each conditional law is an admissible coordinate tilt. Therefore

$$|\mathbb{E}_Q D_i| \leq \mathbb{E}_{Q_{<i}} \left| \mathbb{E}_{Q_i(\cdot \mid X_{<i})} D_i(X_{<i}, \cdot) \right| \leq a_i \Theta_{P_i,d_i}(\rho_i).$$

Sum $h - \mathbb{E}_P h = \sum_i D_i$. For unbounded Lipschitz h, its bounded clippings converge in $L^1(P)$ and $L^1(Q)$ by the first-moment assumptions, preserving the same bound. Duality gives (17); taking $\rho_i = \eta a_i$ proves the tensorization upper bound.

For the reverse inequality, choose admissible coordinate tilts Q_i with score sensitivities ηa_i. Their product is admissible and its summed score is η-Lipschitz under d_a. Moreover,

$$W_{1,d_a}\left(\bigotimes_i Q_i, \bigotimes_i P_i \right) = \sum_i a_i W_{1,d_i}(Q_i, P_i).$$

Every coupling has coordinate costs at least the corresponding Wasserstein distances, and products of optimal coordinate couplings attain their sum. Taking coordinate tilts approaching each supremum proves the reverse bound. If a coordinate profile is infinite, make its transport cost arbitrarily large while leaving all other coordinates untilted. □

The weight a_i enters through both score sensitivity and transport cost; the posterior need not be independent. Appendix A gives a detailed walkthrough and an alternative coupling proof, which motivates Section 6.

5.2 BOUNDED COORDINATES: THE STANDARD BENCHMARK

This benchmark combines a one-coordinate bound with sequential tensorization and tilt integration. If $\log(dQ/dP)$ has oscillation at most ϵ, then $\|Q - P\|_{\mathrm{TV}} \leq \tanh(\epsilon/4)$. On finite or countable spaces, this bounds the discrete-metric profile since W_1 equals total variation. Write $\mathrm{osc}_i(f)$ for the largest change when only coordinate i changes.

Proposition 2 (Bounded-coordinate drift and concentration)**.** *Let $P = \bigotimes_{i=1}^n P_i$ on standard Borel spaces, and let real-valued measurable f, g satisfy $\mathrm{osc}_i(f) \leq c_i$ and $\mathrm{osc}_i(g) \leq \epsilon_i$ for finite $c_i, \epsilon_i \geq 0$. Set $B = \sum_i c_i \tanh(\epsilon_i/4)$ and $v = \sum_i c_i^2$. Then, for $\lambda \geq 0$,*

$$|\mathbb{E}_{Q_g} f - \mathbb{E}_P f| \leq B, \qquad \log \mathbb{E}_{Q_g} e^{\lambda(f - \mathbb{E}_P f)} \leq \lambda B + \lambda^2 v/8. \tag{18}$$

For $v, t > 0$, it follows that $Q_g[|f - \mathbb{E}_P f| \geq B + t] \leq 2e^{-2t^2/v}$.

Proof idea (full proof in Appendix B). On finite or countable coordinates, (17) with weights c_i and conditional budgets ϵ_i gives B; zero weights follow by a limit. Appendix B extends the same averaging argument to general standard Borel spaces. Along $g + uf$, the budgets become $\epsilon_i + uc_i$. Integrate these drifts as in Theorem 2 and use $\tanh'(s) \le 1$ to obtain the MGF bound. At $g = 0$ this recovers McDiarmid's inequality (McDiarmid, 1989).

Learning example. Suppose a pure ϵ-DP transcript under record replacement selects a predictor, evaluated by a loss in $[0, 1]$ on n i.i.d. observations. Its empirical risk $\widehat{R}$ has $c_i = 1/n$ and fresh-data mean R. For almost every transcript and $0 < \delta < 1$,

$$|\widehat{R} - R| \le \underbrace{\tanh(\epsilon/4)}_{\text{selection-bias allowance}} + \underbrace{\sqrt{\frac{\log(2/\delta)}{2n}}}_{\text{sampling term}} \quad \text{with conditional probability at least } 1 - \delta. \tag{19}$$

Thus the familiar sampling scale survives stable selection. For multiple bounded statistical queries, the same drift gives $B_j = \tanh(\epsilon/4)$ in Theorem 1. The full transcript determining the predictor or query must satisfy DP.

5.3 Sampling without replacement

Let P be uniform on the K-subsets of $[N]$, where $1 \le K \le N-1$, with swap distance $d_{\mathrm{sw}}(S, T) = K - |S \cap T|$ and $m = \min(K, N - K)$. The reference law is dependent, but only the profile calculation changes.

Proposition 3 (Slice profile)**.** *For all $\eta \ge 0$,*

$$\Theta_{P,d_{\mathrm{sw}}}(\eta) \le m \tanh(\eta/4). \tag{20}$$

For $\mathrm{Lip}(g) \le \eta$, $\mathrm{Lip}(f) \le L$, and $L, t > 0$, this gives

$$Q_g[|f - \mathbb{E}_P f| \ge mL \tanh(\eta/4) + t] \le 2e^{-2t^2/(mL^2)}. \tag{21}$$

Proof idea (full proof in Appendix C). Reveal the sample in an order and couple completions by swapping labels; passing to the complement gives the smaller factor m. The bound applies conditionally when a stable transcript selects the statistic. Appendix C illustrates this with finite-population averages and adaptively chosen graph cuts.

5.4 Unbounded data: Gaussian and Gamma profiles

Here the metric measures numerical changes, so metric privacy can supply likelihood stability without bounded data or queries. For example, a mean plus Laplace noise is private under ℓ_1 distance on an unbounded domain, although it has no finite replacement-DP guarantee.

Gaussian data. For $P = N(0, \sigma^2 I)$ with $\sigma > 0$,

$$\Theta_{P,\|\cdot\|_2}(\eta) = \sigma^2 \eta, \qquad \eta \ge 0. \tag{22}$$

The entropy argument in Appendix E gives the upper bound, and a linear score shifts the mean by $\sigma^2\eta$ and attains it. Corollary 1 therefore gives bias allowance $\sigma^2 L\eta$ and fluctuation scale σL, even for nonlinear likelihoods whose posterior is not Gaussian. More generally, a smooth density proportional to e^{-V} on $\mathbb{R}^d$ with $\nabla^2 V \succeq \kappa I$ and $\kappa > 0$ satisfies $\Theta_{P,\|\cdot\|_2}(\eta) \le \eta/\kappa$.

Gamma data and changes in fluctuation scale.

Proposition 4 (Gamma profile)**.** *For shape $\alpha > 0$, rate $\beta > 0$, and metric $|x - y|$,*

$$\Theta_{\mathrm{Gamma}(\alpha,\beta),|\cdot|}(\eta) = \frac{\alpha\eta}{\beta(\beta - \eta)}, \qquad 0 \le \eta < \beta, \tag{23}$$

and the profile is infinite for $\eta \ge \beta$.

Proof idea (full proof in Appendix D). The score constraint places the tilt between Gamma laws of rates $\beta + \eta$ and $\beta - \eta$ in stochastic order. Comparing quantiles bounds its Wasserstein distance; $g(x) = \eta x$ attains the upper bound. Its mean diverges as $\eta \uparrow \beta$, ruling out any finite global linear profile bound.

Now take n i.i.d. $\mathrm{Gamma}(\alpha, \beta)$ observations. Let a transcript have an η-Lipschitz log-likelihood under ℓ_1 distance, $0 \le \eta < \beta$, and select a $1/n$-Lipschitz statistic f_π, such as an average of a selected 1-Lipschitz loss. Tensorization and integration give, for almost every transcript and $0 \le \lambda < n(\beta - \eta)$,

$$\begin{aligned}\log \mathbb{E}\Big[e^{\lambda(f_\pi(X) - \mathbb{E}_P f_\pi)} \mid \Pi = \pi\Big] &\le n\alpha \left[-\log\left(1 - \frac{\lambda}{n(\beta-\eta)}\right) - \frac{\lambda}{n\beta}\right] \\ &\le \lambda B_\eta + \frac{V_\eta \lambda^2}{2(1 - b_\eta \lambda)},\end{aligned} \tag{24}$$

with conditional Bernstein parameters

$$B_\eta = \frac{\alpha\eta}{\beta(\beta-\eta)}, \qquad V_\eta = \frac{\alpha}{n(\beta-\eta)^2}, \qquad b_\eta = \frac{1}{n(\beta-\eta)}. \tag{25}$$

For almost every transcript and $t > 0$,

$$\mathbb{P}[|f_\pi(X) - \mathbb{E}_P f_\pi| \ge B_\eta + t \mid \Pi = \pi] \le 2\exp\left(-\frac{t^2}{2(V_\eta + b_\eta t)}\right). \tag{26}$$

Selection changes both the bias allowance B_η and the fluctuation parameters. Relative to $\eta = 0$, V_η grows by $(1 - \eta/\beta)^{-2}$ and b_η by $(1 - \eta/\beta)^{-1}$. Within the admissible tilt class, $g(x) = \eta \sum_i x_i$ and the sample mean attain the mean shift B_η and variance V_η: the tilted coordinates have rate $\beta - \eta$. This is an extremizer of the drift calculation, not a specified selection mechanism; the Bernstein relaxation need not be exact. Appendix D gives the heterogeneous-product formulas and verifies attainment. The full nonlinear profile thus captures an effect absent from the Gaussian and bounded-coordinate bounds.

6 One-sided local profiles and adaptive concentration

Coordinate influences can be small jointly even when their separate worst-case bounds are large. We capture this with dataset-dependent weights. Let $P = \bigotimes_{i=1}^n P_i$ on a finite product of standard Borel spaces $\Omega = \prod_i \Omega_i$.

Definition 2 (One-sided local certificate and profile)**.** *For $L \ge 0$, a measurable real-valued h belongs to $\mathcal{C}_L$ if a measurable field $a : \Omega \to \mathbb{R}_+^n$ satisfies*

$$h(x) - h(y) \le \sum_{i=1}^n a_i(x) \mathbf{1}_{\{x_i \ne y_i\}} \quad \textit{for all } x, y \in \Omega, \qquad \sup_x \|a(x)\|_2 \le L. \tag{27}$$

For $\eta \ge 0$, define the one-sided local posterior drift profile by

$$\Theta_P^{\mathrm{loc},+}(\eta) := \sup_{g \in \mathcal{C}_\eta} \sup_{h \in \mathcal{C}_1} \big(\mathbb{E}_{Q_g} h - \mathbb{E}_P h\big). \tag{28}$$

The weights are *local* to x, but the certificate must hold for every pair, not only single-coordinate changes. Its directed cost need not be a metric. Score and statistic may use different fields, and the profile controls only an upper expectation shift. The certificate implies $\mathrm{osc}(h) \le \sqrt{n} L$, so the tilts and exponential moments are finite.

Theorem 4 (Local posterior drift)**.** *For every product reference law as above,*

$$\Theta_P^{\mathrm{loc},+}(\eta) \le \eta. \qquad \textit{Consequently,} \qquad \mathbb{E}_{Q_g} h - \mathbb{E}_P h \le L\eta \quad (g \in \mathcal{C}_\eta,\ h \in \mathcal{C}_L). \tag{29}$$

Proof idea (full proof in Appendix I). Couple posterior and prior sequentially: the score bounds mismatch probabilities, the statistic bounds their effect, and Cauchy–Schwarz gives $L\eta$. Coordinate transport costs yield the corresponding tensorization argument.

Adding certificates gives $g + uf \in \mathcal{C}_{\eta + uL}$ for $u \ge 0$. Integrating drift and applying Chernoff yields the following corollary; Appendix I gives the full calculation.

Corollary 3 (Concentration under a local likelihood certificate). *If $g \in \mathcal{C}_\eta$ and $f \in \mathcal{C}_L$, then for every $\lambda \geq 0$,*

$$\log \mathbb{E}_{Q_g} e^{\lambda(f - \mathbb{E}_P f)} \leq \lambda L\eta + \frac{\lambda^2 L^2}{2}. \tag{30}$$

For $L > 0$ and $t > 0$,

$$Q_g[f - \mathbb{E}_P f \geq L\eta + t] \leq \exp\left(-\frac{t^2}{2L^2}\right). \tag{31}$$

Taking $g = 0$ gives ordinary concentration. Use the likelihood/Bayes setup of Section 3, replacing the metric constraint by a local certificate. The same bound then holds conditional on almost every π whenever $g_\pi \in \mathcal{C}_{\eta(\pi)}$ and $f_\pi \in \mathcal{C}_{L_\pi}$, with measurable certificates and bounds. The bias allowance is $L_\pi \eta(\pi)$ and the fluctuation scale is L_π. Lower tails require a separate certificate for $-f_\pi$.

Example. On $[0,1]^n$, $f(x) = \|x\|_2$ admits $a(x) = x/\|x\|_2$ for $x \neq 0$ and $a(0) = 0$, so $L = 1$. The upper-tail bound beyond η is $e^{-t^2/2}$ for $g \in \mathcal{C}_\eta$, whereas global bounded differences gives $e^{-2t^2/n}$ at $g = 0$. The fluctuation scale is independent of n; the likelihood budget η may still grow with n. Appendix I gives the derivation and the connection to Talagrand's convex distance (Talagrand, 1995; Boucheron et al., 2013).

7 Connections to entropy and transportation methods

Entropy and transport inequalities provide further ways to estimate the profile of the original law. Let Γ measure squared local sensitivity, for example $\Gamma(u) = \|\nabla u\|_2^2$, with seminorm $\|u\|_\Gamma = \|\sqrt{\Gamma(u)}\|_\infty$. If, for $\rho > 0$, an exponential entropy inequality $\mathrm{Ent}_P(e^u) \leq (2\rho)^{-1}\mathbb{E}_P[e^u \Gamma(u)]$ holds along the required tilt paths, then score and statistic seminorms at most η and L give drift at most $L\eta/\rho$. Under compatibility with the metric and validity for all required Lipschitz scores and tests, this yields $\Theta_{P,d}(\eta) \leq \eta/\rho$. Proposition 6 in Appendix E states the precise hypotheses, including those for restricted inequalities. The proof uses base-law concentration and the tilt entropy, without a posterior entropy inequality.

For $C > 0$, write $T_1(C)$ for $W_{1,d}(R,P)^2 \leq 2C D_{\mathrm{KL}}(R\|P)$ for every $R \ll P$ with finite first moment.

Proposition 5 (Comparison with T_1). *A bound $\Theta_{P,d}(\eta) \leq C\eta$ for every $\eta \geq 0$ implies $T_1(C)$. Conversely, $T_1(C)$ implies $\Theta_{P,d}(\eta) \leq 2C\eta$. The converse coefficient two is optimal uniformly over reference laws and metrics.*

The complete proof and sharpness examples are in Appendix F. The forward direction uses the log-Laplace characterization of T_1 (Bobkov & Götze, 1999); the converse uses the symmetric entropy identity. Linear profiles thus describe the usual sub-Gaussian regime up to a sharp factor, while keeping likelihood sensitivity explicit for transfer and adaptive concentration. Nonlinear profiles retain information beyond this regime.

8 Conclusion

Posterior drift profiles provide a common tool for accuracy transfer and concentration. For transfer, they separate the sampling law and geometry from a short conditional-expectation argument, with the profile supplying the selection-bias allowance. Building on existing posterior proofs (Jung et al., 2020), the framework extends standard DP transfer under Hamming distance to pure metric privacy under our likelihood assumptions.

The framework also provides a concentration method alongside entropy and transportation methods: bound expectation drift along exponential tilts and integrate. The same profile of the original law yields ordinary concentration and concentration after stable selection, without a new functional inequality for the posterior. Exact tensorization and explicit examples make the calculation reusable across sampling laws and geometries; nonlinear profiles describe how selection can enlarge both bias and fluctuation bounds.

REFERENCES

Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence II: Optimal sampling and concentration via restricted modified log-Sobolev inequalities. *arXiv preprint arXiv:2111.03247*, 2021. URL https://arxiv.org/abs/2111.03247. Detailed Part II manuscript; merged conference version published at STOC 2022.

Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence: Optimal mixing of down-up random walks. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1418–1430. ACM, 2022. doi: 10.1145/3519935.3520048.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In Jacques Azéma and Marc Yor (eds.), *Séminaire de Probabilités XIX 1983/84*, volume 1123 of *Lecture Notes in Mathematics*, pp. 177–206. Springer, 1985. doi: 10.1007/BFb0075847.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. URL https://jmlr.org/papers/v3/bartlett02a.html.

Raef Bassily and Yoav Freund. Typical stability. *arXiv preprint arXiv:1604.03336*, 2016. URL https://arxiv.org/abs/1604.03336.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pp. 1046–1059. ACM, 2016. doi: 10.1145/2897518.2897566. URL https://arxiv.org/abs/1511.02513.

Daniel Berend and Aryeh Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18:1–7, 2013. doi: 10.1214/ECP.v18-2359. URL https://arxiv.org/abs/1210.3248. Paper No. 3.

Sergey G. Bobkov and Friedrich Götze. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 1999. doi: 10.1006/jfan.1998.3326.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002. URL https://jmlr.org/papers/v2/bousquet02a.html.

Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *Privacy Enhancing Technologies*, volume 7981 of *Lecture Notes in Computer Science*, pp. 82–102. Springer, 2013. doi: 10.1007/978-3-642-39077-7_5.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, volume 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006. doi: 10.1007/11681878_14.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pp. 117–126. ACM, 2015. doi: 10.1145/2746539.2746580. URL https://arxiv.org/abs/1411.2664.

Amedeo Roberto Esposito and Michael Gastpar. From generalisation error to transportation-cost inequalities and back. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 294–299, 2022. doi: 10.1109/ISIT50566.2022.9834354. URL https://arxiv.org/abs/2202.03956.

Max Fathi, Dan Mikulincer, and Yair Shenfeld. Transportation onto log-Lipschitz perturbations. *Calculus of Variations and Partial Differential Equations*, 63(3):61, 2024. doi: 10.1007/s00526-023-02652-x. URL https://arxiv.org/abs/2305.03786.

Leonard Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975. doi: 10.2307/2373688.

Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3):824–839, 2020. doi: 10.1109/JSAIT.2020.3040992. URL https://arxiv.org/abs/2005.08044v6. Corrected author version, 2023.

Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, volume 151 of *Leibniz International Proceedings in Informatics*, pp. 31:1–31:17. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020. doi: 10.4230/LIPIcs.ITCS.2020.31. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2020.31.

Ksenia A. Khudiakova, Jan Maas, and Francesco Pedrotti. L^∞-optimal transport of anisotropic log-concave measures and exponential convergence in Fisher's infinitesimal model. *The Annals of Applied Probability*, 35(3):1913–1940, 2025. doi: 10.1214/25-AAP2162. URL https://arxiv.org/abs/2402.04151.

Aryeh Kontorovich, Menachem Sadigurschi, and Uri Stemmer. Adaptive data analysis with correlated observations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11483–11498. PMLR, 2022. URL https://proceedings.mlr.press/v162/kontorovich22a.html.

Christophe Ley, Gesine Reinert, and Yvik Swan. Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. *The Annals of Applied Probability*, 27(1):216–241, 2017. doi: 10.1214/16-AAP1202. URL https://arxiv.org/abs/1510.05826.

Adrian Tovar Lopez and Varun Jog. Generalization error bounds using Wasserstein distances. In *2018 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2018. doi: 10.1109/ITW.2018.8613445.

Katalin Marton. A measure concentration inequality for contracting Markov chains. *Geometric and Functional Analysis*, 6(3):556–571, 1996. doi: 10.1007/BF02249263. Erratum: 7:609–613, 1997, doi:10.1007/s000390050021.

David A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999. doi: 10.1023/A:1007618624809.

Colin McDiarmid. On the method of bounded differences. In J. Siemons (ed.), *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

Kobbi Nissim and Uri Stemmer. Concentration bounds for high sensitivity functions through differential privacy. *Journal of Privacy and Confidentiality*, 9(1), 2019. doi: 10.29012/jpc.658.

Borja Rodríguez-Gálvez, Germán Bassi, Ragnar Thobaben, and Mikael Skoglund. Tighter expected generalization error bounds via Wasserstein distance. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19109–19121. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9f975093da0252e2c0ae181d74c90dc6-Abstract.html.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 1232–1240. PMLR, 2016. URL https://proceedings.mlr.press/v51/russo16.html.

Moshe Shenfeld and Katrina Ligett. A necessary and sufficient stability notion for adaptive generalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c5df4f4eabf1cbcfeb50fbbf97c5289f-Abstract.html.

Moshe Shenfeld and Katrina Ligett. Generalization in the face of adaptivity: A Bayesian perspective. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52153–52165. Curran Associates, Inc., 2023. doi: 10.52202/075280-2271. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3c01875a052f81d27a5211df096cd91-Abstract-Conference.html.

Thomas Steinke and Jonathan Ullman. Subgaussian tail bounds via stability arguments. *arXiv preprint arXiv:1701.03493*, 2017. URL https://arxiv.org/abs/1701.03493.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3437–3452. PMLR, 2020. URL https://proceedings.mlr.press/v125/steinke20a.html.

Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'IHÉS*, 81:73–205, 1995. doi: 10.1007/BF02699376. URL https://www.numdam.org/item/PMIHES_1995__81__73_0/.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. doi: 10.1007/978-3-540-71050-9.

Olivier Wintenberger. Weak transport inequalities and applications to exponential and oracle inequalities. *Electronic Journal of Probability*, 20(114):1–27, 2015. doi: 10.1214/EJP.v20-3558. URL https://arxiv.org/abs/1207.4951.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, volume 30, pp. 2524–2533, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/ad71c82b22f4f65b9398f76d8be4c615-Abstract.html.

A EXACT PRODUCT TENSORIZATION

This appendix expands the complete proof in Section 5.1 and then gives a second proof by sequential coupling. The upper bound must hold for every admissible posterior, including dependent ones. We first show that conditioning preserves the relevant coordinate score bounds, and then add the resulting coordinate drifts. To obtain *equality*, we also construct product tilts approaching the upper bound. The coupling argument provides an alternative upper bound and uses the same matching construction.

Throughout, $P = \bigotimes_{i=1}^{n} P_i$ and $d_a(x, y) = \sum_i a_i d_i(x_i, y_i)$, where $a_i > 0$ and each P_i has finite first moment on a Polish metric space (Ω_i, d_i). Write $x_{<i} = (x_1, \dots, x_{i-1})$ for a prefix, $x_{>i}$ for the remaining coordinates, and $P_{>i} = \bigotimes_{j>i} P_j$. For a law Q, $Q_{<i}$ denotes its prefix marginal and $Q_i(\cdot \mid x_{<i})$ its conditional law of coordinate i. Only P is assumed to be a product law. All posteriors considered below have finite first moment; profile values may nevertheless be infinite.

A.1 WHY CONDITIONAL SCORES RETAIN THE SENSITIVITY BOUND

Let $Q = Q_g$ be admissible with $\mathrm{Lip}_{d_a}(g) \le \eta$, and put $r = dQ/dP = e^g/Z_g$. Normalization subtracts a constant from the score, so $\log r$ has the same Lipschitz bound. To find the conditional density of coordinate i, first integrate out the unrevealed coordinates:

$$W_i(x_{<i}, u) = \int r(x_{<i}, u, z_{>i})\, P_{>i}(dz_{>i}), \qquad Z_i(x_{<i}) = \int W_i(x_{<i}, v)\, P_i(dv).$$

Here Z_i is the density of the prefix marginal $Q_{<i}$ relative to $P_{<i}$. Thus, for $Q_{<i}$-almost every prefix, $0 < Z_i(x_{<i}) < \infty$ and the conditional density is

$$\frac{dQ_i(\cdot \mid x_{<i})}{dP_i}(u) = \frac{W_i(x_{<i}, u)}{Z_i(x_{<i})}.$$

This is simply the joint density of the prefix and current coordinate divided by the prefix density.

Changing only coordinate i changes g by at most $\eta a_i d_i(u, v)$. Consequently,

$$r(x_{<i}, u, z_{>i}) \le e^{\eta a_i d_i(u,v)} r(x_{<i}, v, z_{>i}) \quad \Longrightarrow \quad W_i(x_{<i}, u) \le e^{\eta a_i d_i(u,v)} W_i(x_{<i}, v).$$

The implication follows by integrating the same pointwise inequality over the prior suffix. For a prefix with $0 < Z_i < \infty$, some coordinate value has $0 < W_i < \infty$; comparison in both directions then gives this property at every point of the coordinate support. Taking logarithms and interchanging u, v yields

$$\left| \log \frac{W_i(x_{<i}, u)}{Z_i(x_{<i})} - \log \frac{W_i(x_{<i}, v)}{Z_i(x_{<i})} \right| \le \eta a_i d_i(u, v).$$

The denominator cancels because the prefix is fixed. Each conditional law is therefore a tilt of P_i with score sensitivity at most ηa_i.

These conditional tilts also satisfy the required moment condition. For a fixed reference point o_i in the coordinate support,

$$\int \left[\int d_i(u, o_i)\, Q_i(du \mid x_{<i}) \right] Q_{<i}(dx_{<i}) = \mathbb{E}_Q d_i(X_i, o_i) < \infty.$$

The last inequality follows from the first moment of Q under d_a and $a_i > 0$. The inner integral is thus finite for almost every posterior prefix, so the coordinate profiles apply.

A.2 DETAILED UPPER BOUND BY SUCCESSIVE AVERAGING

We prove the more general sequential bound (17). Assume $Q \ll P$ has finite first moment and each conditional score has Lipschitz constant at most a deterministic ρ_i. The preceding subsection gives the special choice $\rho_i = \eta a_i$ when $Q = Q_g$.

Step 1: decompose a test function into coordinate increments. Take a bounded function h with $\mathrm{Lip}_{d_a}(h) \le 1$ and define

$$H_i(x_{\le i}) = \int h(x_{\le i}, z_{>i})\, P_{>i}(dz_{>i}), \qquad H_0 = \mathbb{E}_P h, \qquad H_n = h.$$

The function H_i keeps the first i coordinates fixed and averages the rest under their original independent laws. Set

$$D_i(x_{\le i}) = H_i(x_{\le i}) - H_{i-1}(x_{<i}). \qquad \text{Then} \qquad h(x) - \mathbb{E}_P h = \sum_{i=1}^{n} D_i(x_{\le i}).$$

This identity is just telescoping. It remains true when x is sampled from Q, even though the averages defining H_i use P.

Step 2: check centering and sensitivity under the coordinate prior. Independence under P gives

$$\int H_i(x_{<i}, u)\, P_i(du) = H_{i-1}(x_{<i}), \qquad \int D_i(x_{<i}, u)\, P_i(du) = 0.$$

For fixed $x_{<i}$, subtracting $H_{i-1}(x_{<i})$ does not change sensitivity in the current coordinate. Hence

$$\begin{aligned} |D_i(x_{<i}, u) - D_i(x_{<i}, v)| &= |H_i(x_{<i}, u) - H_i(x_{<i}, v)| \\ &\le \int |h(x_{<i}, u, z_{>i}) - h(x_{<i}, v, z_{>i})|\, P_{>i}(dz_{>i}) \\ &\le a_i d_i(u, v). \end{aligned}$$

Thus $D_i(x_{<i}, \cdot)$ is a centered, a_i-Lipschitz coordinate statistic. These are the prior Doob increments, but no martingale result is needed: the two displayed identities are the only properties we use. In particular, the increments need not have mean zero under the posterior.

Step 3: bound the posterior shift of each increment and add. For almost every posterior prefix, the conditional first moment is finite by the same disintegration argument as above. The definition of the coordinate profile and the centering from Step 2 therefore imply

$$\begin{aligned} \left|\mathbb{E}_{Q_i(\cdot|x_{<i})} D_i(x_{<i}, X_i)\right| &= \left|\mathbb{E}_{Q_i(\cdot|x_{<i})} D_i(x_{<i}, X_i) - \mathbb{E}_{P_i} D_i(x_{<i}, X_i)\right| \\ &\le a_i \Theta_{P_i, d_i}(\rho_i). \end{aligned}$$

Now average over the prefix under $Q_{<i}$. The tower property and the triangle inequality give

$$|\mathbb{E}_Q D_i| \le \mathbb{E}_{Q_{<i}} \left|\mathbb{E}_{Q_i(\cdot|X_{<i})} D_i(X_{<i}, X_i)\right| \le a_i \Theta_{P_i, d_i}(\rho_i).$$

Taking expectations in the telescoping identity and summing proves

$$|\mathbb{E}_Q h - \mathbb{E}_P h| \le \sum_{i=1}^{n} |\mathbb{E}_Q D_i| \le \sum_{i=1}^{n} a_i \Theta_{P_i, d_i}(\rho_i). \tag{32}$$

Posterior dependence causes no difficulty: the bound is applied conditionally at each prefix before averaging over that prefix.

Step 4: pass from test functions to Wasserstein distance. For an unbounded 1-Lipschitz h, subtract $h(o)$ at a fixed point o and clip the resulting function at levels $\pm R$. The clipped functions remain 1-Lipschitz and are bounded in absolute value by $|h(x) - h(o)| \le d_a(x, o)$. Finite first moments under both P and Q give convergence in $L^1(P)$ and $L^1(Q)$. Thus (32) holds for every 1-Lipschitz test. Kantorovich–Rubinstein duality says that the supremum of these expectation differences is precisely Wasserstein distance:

$$W_{1,d_a}(Q, P) = \sup_{\mathrm{Lip}_{d_a}(h) \le 1} |\mathbb{E}_Q h - \mathbb{E}_P h| \le \sum_i a_i \Theta_{P_i, d_i}(\rho_i).$$

This proves the sequential form of Theorem 3. For an admissible joint tilt, substitute $\rho_i = \eta a_i$ and then take the supremum over all scores of sensitivity at most η:

$$\Theta_{P, d_a}(\eta) \le \sum_i a_i \Theta_{P_i, d_i}(\eta a_i).$$

If a profile on the right is infinite, the upper bound is automatic; the lower-bound argument below still establishes equality.

A.3 WHY THE UPPER BOUND IS EXACT

The upper bound permits dependent posteriors. To approach it, however, independent coordinate tilts suffice. Choose admissible $Q_i = e^{g_i} P_i / Z_{g_i}$ with $\mathrm{Lip}_{d_i}(g_i) \le \eta a_i$ and set $g(x) = \sum_i g_i(x_i)$. Then

$$|g(x) - g(y)| \le \sum_i \eta a_i d_i(x_i, y_i) = \eta d_a(x, y).$$

Independence under P gives $Z_g = \prod_i Z_{g_i}$, and therefore $Q_g = \bigotimes_i Q_i$. This joint tilt is admissible because each coordinate tilt has finite first moment and there are finitely many coordinates.

For these product laws, the additive metric gives the exact identity

$$W_{1,d_a}\left(\bigotimes_i Q_i, \bigotimes_i P_i\right) = \sum_i a_i W_{1,d_i}(Q_i, P_i).$$

To see both inequalities, take any coupling of the two product laws. Its ith coordinate pair is a coupling of Q_i and P_i, so its expected cost is at least the sum on the right. Conversely, take an optimal coupling for each coordinate pair and generate those pairs independently. This is a coupling of the product laws whose cost is exactly that sum.

Profiles are suprema, so their maximizing tilts need not exist. If all coordinate profile values are finite, fix $\varepsilon > 0$ and choose the ith tilt so that

$$W_{1,d_i}(Q_i, P_i) \ge \Theta_{P_i,d_i}(\eta a_i) - \frac{\varepsilon}{n a_i}.$$

The resulting joint tilt has transport cost at least $\sum_i a_i \Theta_{P_i,d_i}(\eta a_i) - \varepsilon$. Letting $\varepsilon \downarrow 0$ gives the reverse inequality. If instead $\Theta_{P_k,d_k}(\eta a_k) = \infty$ for some k, choose an admissible tilt in that coordinate with cost greater than M/a_k, and leave all other coordinates untilted. The joint transport cost exceeds M for every prescribed $M > 0$. This proves the exact identity, including infinite profile values.

A.4 ALTERNATIVE PROOF BY SEQUENTIAL COUPLING

The preceding upper-bound proof controls expectation differences for every Lipschitz test function. We can instead construct a pair (X, Y) with laws Q and P and directly bound the distance between them. The matching lower bound remains the one in Appendix A.3.

The useful intermediate statement is valid for any Q with finite first moment, without a likelihood-sensitivity assumption:

$$W_{1,d_a}(Q, P) \le \sum_{i=1}^{n} a_i \, \mathbb{E}_{Q_{<i}} W_{1,d_i}(Q_i(\cdot \mid X_{<i}), P_i)\,. \tag{33}$$

It compares the full joint laws by comparing one posterior conditional law at a time with its prior coordinate law.

Step 1: construct the coordinate pairs. For each i and almost every posterior prefix $x_{<i}$, choose an optimal coupling $\gamma_i(x_{<i})$ of $Q_i(\cdot \mid x_{<i})$ and P_i. The choices can be made as measurable kernels: for Polish spaces, the sets of optimal couplings are nonempty and compact with a Borel graph, so measurable selection applies (Villani, 2009, Corollary 5.22). Conditional first moments are finite almost surely; set the kernels arbitrarily on exceptional prefixes. Starting at $i = 1$, generate pairs according to

$$\mathcal{L}(X_i, Y_i \mid X_{<i}, Y_{<i}) = \gamma_i(X_{<i}).$$

The choice depends on the previously generated X coordinates, but does not use the previous Y coordinates.

Step 2: verify that the full marginals are Q and P. By the first marginal of each coordinate coupling,

$$\mathcal{L}(X_i \mid X_{<i}, Y_{<i}) = Q_i(\cdot \mid X_{<i}).$$

Ignoring the Y history recovers the conditional laws defining Q, so successive generation gives $X \sim Q$. The second marginal gives

$$\mathcal{L}(Y_i \mid X_{<i}, Y_{<i}) = P_i.$$

In particular, $\mathcal{L}(Y_i \mid Y_{<i}) = P_i$: the conditional law does not vary with the preceding coordinates. Hence the Y_i are independent with laws P_i, and $Y \sim P$. This is why the construction works even when the X_i are dependent.

Step 3: add the transport costs. Optimality of the coordinate couplings and the tower property give

$$\begin{aligned}\mathbb{E}d_a(X,Y) &= \sum_i a_i \, \mathbb{E}_{Q_{<i}} \int d_i(u,v)\, \gamma_i(X_{<i})(du,dv) \\ &= \sum_i a_i \, \mathbb{E}_{Q_{<i}} W_{1,d_i}(Q_i(\cdot \mid X_{<i}), P_i)\,.\end{aligned}$$

Wasserstein distance is the infimum of expected costs over all joint couplings, so this particular construction proves (33). If the conditional scores have budgets ρ_i, their coordinate Wasserstein distances are at most $\Theta_{P_i,d_i}(\rho_i)$. Substitution recovers (17). Appendix A.1 supplies $\rho_i = \eta a_i$ for a joint score of sensitivity η, and Appendix A.3 completes exact tensorization.

Connection to the one-sided local extension. For the additive metric, only the expected coordinate distances enter the final bound. The local certificates of Section 6 instead have weights depending on the realized dataset. The sequential construction can retain the information needed for those weights: use maximal coordinate couplings and control the mismatch probabilities $\Pr(Y_i \neq X_i \mid X)$. Appendix I proves those conditional bounds and pairs them with the local certificates. Thus the coupling proof explains the connection to the asymmetric extension, while the averaging proof gives a direct route to the metric tensorization identity.

B Bounded coordinates and likelihood oscillations

We prove Proposition 2 by combining a one-coordinate estimate with the sequential bound of Theorem 3, then integrating along further tilts as in Theorem 2. We also explain how the same averaging argument covers general standard Borel coordinate spaces.

Step 1: the one-coordinate bound. Let $Q \ll P$ have density r with $\operatorname{ess\,sup} \log r - \operatorname{ess\,inf} \log r \leq \epsilon$. Write $m = \operatorname{ess\,inf} r$ and $M = \operatorname{ess\,sup} r$. Then $0 < m \leq 1 \leq M < \infty$ and $M/m \leq e^{\epsilon}$. If $m = M$, then $Q = P$. Otherwise the chord bound for the convex function $z \mapsto |z-1|$, together with $\mathbb{E}_P r = 1$, gives

$$\|Q - P\|_{\mathrm{TV}} = \tfrac{1}{2}\mathbb{E}_P|r-1| \leq \frac{(1-m)(M-1)}{M-m} \leq \frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}} \leq \tanh(\epsilon/4).$$

The middle inequality is equivalent to $2\sqrt{mM} \leq 1+mM$. Thus a function with oscillation at most c has expectation drift at most $c\tanh(\epsilon/4)$. In particular, if the underlying metric has diameter at most one, every 1-Lipschitz function has this oscillation bound, which also proves $\Theta_{P,d}(\epsilon) \leq \tanh(\epsilon/4)$.

Step 2: apply sequential tensorization. Now let $P = \bigotimes_i P_i$, let $Q_g \propto e^g P$, and assume the global coordinate bounds $\mathrm{osc}_i(g) \leq \epsilon_i$ and $\mathrm{osc}_i(f) \leq c_i$. Finite coordinate oscillations bound the total oscillations of f and g by $\sum_i c_i$ and $\sum_i \epsilon_i$. Thus these functions are bounded and all tilts used below are well-defined.

First suppose each coordinate space is finite or countable and equip it with $d_{\mathrm{disc}}(u,v) = \mathbf{1}_{\{u \neq v\}}$. For this metric, Lipschitz sensitivity is oscillation and Wasserstein distance equals total variation, so Step 1 gives

$$\Theta_{P_i,d_{\mathrm{disc}}}(\epsilon_i) \leq \tanh(\epsilon_i/4).$$

Suffix integration preserves the score bound ϵ_i for each posterior conditional law, by Appendix A.1. For $\tau > 0$, use the weighted Hamming metric $d_{c+\tau}(x,y) = \sum_i (c_i + \tau)\mathbf{1}_{\{x_i \neq y_i\}}$. Changing

coordinates one at a time shows that f is 1-Lipschitz under this metric. The sequential form (17) therefore gives

$$|\mathbb{E}_{Q_g} f - \mathbb{E}_P f| \le W_{1,d_{c+\tau}}(Q_g, P) \le \sum_i (c_i + \tau)\Theta_{P_i,d_{\mathrm{disc}}}(\epsilon_i) \le \sum_i (c_i + \tau)\tanh(\epsilon_i/4).$$

Letting $\tau \downarrow 0$ yields the allowance B, including when some c_i vanish. The separate budgets ϵ_i need not be proportional to c_i, which is why we use the sequential form of the theorem. For a common budget η, exact tensorization also gives the unweighted Hamming profile appearing in Table 1:

$$\Theta_{P,d_H}(\eta) = \sum_i \Theta_{P_i,d_{\mathrm{disc}}}(\eta) \le n\tanh(\eta/4).$$

Extension to general standard Borel spaces. The discrete metric on an uncountable space is not separable, so the Polish-metric statement of Theorem 3 does not directly apply there. Its averaging proof in Appendix A.2, however, uses only coordinate expectation comparisons. Let D_i be the prior averaging increments of f defined there. Averaging the unrevealed coordinates preserves the oscillation bound c_i, and $\mathbb{E}_{P_i} D_i(x_{<i}, X_i) = 0$ for every prefix. Suffix integration likewise preserves the conditional score oscillation bound ϵ_i. Step 1 therefore yields

$$\left|\mathbb{E}_{Q_i(\cdot|x_{<i})} D_i(x_{<i}, X_i)\right| \le c_i \|Q_i(\cdot \mid x_{<i}) - P_i\|_{\mathrm{TV}} \le c_i \tanh(\epsilon_i/4)$$

for almost every posterior prefix. Averaging over $Q_{<i}$ and summing the same telescoping decomposition proves

$$|\mathbb{E}_{Q_g} f - \mathbb{E}_P f| \le B, \qquad B := \sum_i c_i \tanh(\epsilon_i/4). \tag{34}$$

Step 3: integrate the coordinate drifts. For $u \ge 0$, the score $g + uf$ has coordinate oscillations at most $\epsilon_i + uc_i$. Differentiating the centered log-MGF as in Theorem 2 and using (34) for this new score gives

$$\log \mathbb{E}_{Q_g} e^{\lambda(f - \mathbb{E}_P f)} \le \int_0^\lambda \sum_i c_i \tanh\left(\frac{\epsilon_i + uc_i}{4}\right) du \tag{35}$$

$$= 4\sum_i \left[\log\cosh\left(\frac{\epsilon_i + \lambda c_i}{4}\right) - \log\cosh(\epsilon_i/4)\right]. \tag{36}$$

The exponential moments are finite because f and g are bounded. Since $\tanh'(v) \le 1$,

$$\log \mathbb{E}_{Q_g} e^{\lambda(f - \mathbb{E}_P f)} \le \lambda B + \frac{\lambda^2}{8}\sum_i c_i^2.$$

For $v := \sum_i c_i^2 > 0$, Chernoff's bound with $\lambda = 4t/v$ yields

$$Q_g\{f - \mathbb{E}_P f \ge B + t\} \le e^{-2t^2/v}, \qquad t \ge 0.$$

Applying the same argument to $-f$ yields the two-sided bound $Q_g\{|f - \mathbb{E}_P f| \ge B + t\} \le 2e^{-2t^2/v}$. If $v = 0$, f is constant and the conclusion for $t > 0$ is immediate. Global score bounds are used here so that the required conditional bounds remain valid after the additional tilt by e^{uf}.

Statistical queries after private selection. For the familiar statistical-query example, let $X \sim P_0^{\otimes n}$, let the full transcript be pure ϵ-DP under record replacement, and let it select $f_\pi(X) = n^{-1}\sum_i \phi_\pi(X_i)$ with $0 \le \phi_\pi \le 1$. Substituting $c_i = 1/n$ and $\epsilon_i = \epsilon$ gives, for almost every transcript and $t > 0$,

$$\mathbb{P}[|f_\pi(X) - \mathbb{E}_P f_\pi| \ge \tanh(\epsilon/4) + t \mid \Pi = \pi] \le 2e^{-2nt^2}. \tag{37}$$

The usual sampling scale survives, with an explicit allowance for selection. This is the learning benchmark in Section 5.2; its role is to check that the general method retains the familiar bounded-differences guarantee.

C SAMPLING WITHOUT REPLACEMENT

We prove Proposition 3. Let P be uniform on the K-subsets of $[N]$, where $1 \le K \le N-1$, and let $d_{\mathrm{sw}}(S,T) = K - |S \cap T|$. Complementation is an isometry from the K-slice to the $(N-K)$-slice and preserves uniform laws and score bounds. It suffices to treat $K \le N/2$, so that $m = K$.

Let ν be the uniform law of an ordered sample $Z = (Z_1, \dots, Z_K)$ of distinct elements of $[N]$, and write $\pi(z) = \{z_1, \dots, z_K\}$. Given a tilt Q on the slice with density r and $\mathrm{Lip}_{d_{\mathrm{sw}}}(\log r) \le \eta$, lift it to the ordered space by $d\widehat{Q}/d\nu(z) = r(\pi(z))$. For a fixed prefix $z_{<i}$, the prior conditional law ν_i of Z_i is uniform on the remaining labels. The posterior conditional density, relative to ν_i, is proportional to

$$W_i(z_{<i}, u) = \mathbb{E}_\nu[r(\pi(Z)) \mid Z_{<i} = z_{<i},\ Z_i = u].$$

For two remaining labels u, v, transpose u and v in every suffix completion after u. This is a bijection onto the suffix completions after v, and it preserves their uniform probabilities. The two completed unordered sets coincide if the original suffix contains v; otherwise they differ by the single swap $u \mapsto v$. The score bound implies

$$W_i(z_{<i}, u) \le e^{\eta} W_i(z_{<i}, v).$$

Thus every posterior conditional has log-density oscillation at most η relative to its prior conditional.

Let h be L-Lipschitz on the slice, and set

$$H_i(z_{\le i}) = \mathbb{E}_\nu[h(\pi(Z)) \mid Z_{\le i} = z_{\le i}], \qquad D_i = H_i - H_{i-1}, \qquad H_0 = \mathbb{E}_P h.$$

The same suffix bijection shows that, for each prefix, $u \mapsto D_i(z_{<i}, u)$ has oscillation at most L. Its expectation under ν_i is zero. The total-variation estimate in Appendix B therefore gives

$$\left|\mathbb{E}_{\widehat{Q}} D_i(Z_{\le i})\right| \le L \tanh(\eta/4).$$

Summing the K increments proves

$$|\mathbb{E}_Q h - \mathbb{E}_P h| \le mL \tanh(\eta/4), \qquad \Theta_{P,d_{\mathrm{sw}}}(\eta) \le m \tanh(\eta/4).$$

For an L-Lipschitz f and a score g with Lipschitz constant at most η, Theorem 2 now yields

$$\begin{aligned}\log \mathbb{E}_{Q_g} e^{\lambda(f - \mathbb{E}_P f)} &\le \int_0^\lambda mL \tanh\left(\frac{\eta + uL}{4}\right) du \\ &\le \lambda mL \tanh(\eta/4) + \frac{mL^2\lambda^2}{8}.\end{aligned}$$

For $L > 0$, applying Chernoff's bound to f and $-f$ gives

$$Q_g\{|f - \mathbb{E}_P f| \ge mL \tanh(\eta/4) + t\} \le 2\exp\left(-\frac{2t^2}{mL^2}\right), \qquad t \ge 0.$$

The case $L = 0$ is immediate for positive deviations.

Finite-population averages. For $f(S) = K^{-1}\sum_{i \in S} \phi(i)$ with $\phi : [N] \to [0,1]$, the fresh-sample mean is $N^{-1}\sum_i \phi(i)$. Since $L = 1/K$, the bias allowance is $(m/K)\tanh(\eta/4)$ and the upper bound on the two-sided tail is $2e^{-2K^2t^2/m}$. A stable transcript can select ϕ: conditioning fixes this function and reweights the slice law, after which the same profile applies. This illustrates how changing the sampling model changes the drift calculation, while the transfer and concentration arguments stay the same. The present bound uses global sensitivity and does not exploit the variance of a particular population statistic.

Example: a graph cut chosen from the observed graph. For $G \sim G(n, M)$, identify its edges with an M-subset of the $N_e = \binom{n}{2}$ possible edges, and let $m = M \wedge (N_e - M) > 0$. Suppose a pure ϵ-DP transcript under edge swaps selects a cut $U_\pi \subseteq [n]$. Its crossing-edge count $e_G(U_\pi, U_\pi^c)$ changes by at most one under a swap. Holding the cut fixed, its fresh-graph expectation is $p|U_\pi|(n - |U_\pi|)$, where $p = M/N_e$. Thus, for almost every transcript and $t > 0$,

$$\begin{aligned}&\mathbb{P}[|e_G(U_\pi, U_\pi^c) - p|U_\pi|(n - |U_\pi|)| \ge m \tanh(\epsilon/4) + t \mid \Pi = \pi] \\ &\qquad \le 2e^{-2t^2/m}.\end{aligned} \tag{38}$$

Here the reference data are dependent because the total number of edges is fixed, and the cut itself is chosen adaptively.

D GAMMA DRIFT AND CONCENTRATION

We prove Proposition 4 and record its posterior extension. Fix shape $\alpha > 0$ and rate $\beta > 0$. Let P be the corresponding Gamma law, equipped with the metric $|x - y|$. Fix $0 \le \eta < \beta$ and an η-Lipschitz score g. The tilt Q_g is normalizable and has finite first moment because $e^{g(x)}$ is bounded above by a constant times $e^{\eta x}$. Define

$$P^+ = \mathrm{Gamma}(\alpha, \beta - \eta), \qquad P^- = \mathrm{Gamma}(\alpha, \beta + \eta).$$

The likelihood ratio dQ_g/dP^+ is proportional to $e^{g(x)-\eta x}$, which is nonincreasing, whereas dQ_g/dP^- is proportional to $e^{g(x)+\eta x}$, which is nondecreasing. A nonincreasing likelihood ratio decreases every upper-tail probability, since its covariance with an increasing tail indicator under the reference law is nonpositive. The increasing case reverses this conclusion. Hence

$$P^- \preceq_{\mathrm{st}} Q_g \preceq_{\mathrm{st}} P^+.$$

If q is the quantile function of P, the quantiles of $P^\pm$ are $q^\pm(u) = \beta q(u)/(\beta \mp \eta)$. The stochastic sandwich gives $q^-(u) \le q_{Q_g}(u) \le q^+(u)$. Since

$$q(u) - q^-(u) = \frac{\eta}{\beta + \eta} q(u) \le \frac{\eta}{\beta - \eta} q(u) = q^+(u) - q(u),$$

the one-dimensional quantile formula for Wasserstein distance yields

$$W_1(Q_g, P) \le \int_0^1 \frac{\eta}{\beta - \eta} q(u)\, du = \frac{\alpha\eta}{\beta(\beta - \eta)}.$$

Equality is achieved by $g(x) = \eta x$, for which $Q_g = P^+$ and $W_1(P^+, P) = \mathbb{E}_{P^+} X - \mathbb{E}_P X$. Thus

$$\Theta_{P,|\cdot|}(\eta) = \frac{\alpha\eta}{\beta(\beta - \eta)}, \qquad 0 \le \eta < \beta. \tag{39}$$

For completeness, the profile is infinite for $\eta \ge \beta$: admissible linear scores with slopes tending to β from below have unbounded drift, although each individual tilt has finite first moment.

Weighted products and ordinary concentration. Consider independent Gamma coordinates with shapes $\alpha_i > 0$ and rates $\beta_i > 0$, product law P, and metric $d_a(x, y) = \sum_i a_i |x_i - y_i|$ with $a_i > 0$. Set

$$b_i = \frac{a_i}{\beta_i}, \qquad b = \max_i b_i, \qquad V = \sum_i \alpha_i b_i^2.$$

Exact tensorization and (39) imply

$$\Theta_{P,d_a}(\eta) = \sum_i \frac{\alpha_i \eta b_i^2}{1 - \eta b_i}, \qquad 0 \le \eta < 1/b. \tag{40}$$

For a 1-Lipschitz f under d_a, integration along its exponential tilts gives, for $0 \le \lambda < 1/b$,

$$\begin{aligned} \log \mathbb{E}_P e^{\lambda(f - \mathbb{E}_P f)} &\le \sum_i \alpha_i [-\log(1 - \lambda b_i) - \lambda b_i] \\ &\le \frac{\lambda^2 V}{2(1 - \lambda b)}. \end{aligned} \tag{41}$$

Here the last step uses $-\log(1 - z) - z \le z^2/[2(1 - z)]$ for $0 \le z < 1$. The exponential moments exist because $|f(x) - f(0)| \le \sum_i a_i x_i$. Taking $\lambda = t/(V + bt)$ in Chernoff's bound yields

$$P\{f - \mathbb{E}_P f \ge t\} \le \exp\left(-\frac{t^2}{2(V + bt)}\right), \qquad t > 0.$$

The same bound applies to $-f$; the two-sided version has an additional factor of two.

Posterior concentration. Keep f 1-Lipschitz under d_a, let $\mathrm{Lip}_{d_a}(g) \le \eta < 1/b$, and define

$$B_\eta = \sum_i \frac{\alpha_i \eta b_i^2}{1-\eta b_i}, \qquad V_\eta = \sum_i \alpha_i \left(\frac{b_i}{1-\eta b_i}\right)^2, \qquad b_\eta = \max_i \frac{b_i}{1-\eta b_i}. \tag{42}$$

For $0 \le \lambda < 1/b_\eta$, both g and $g+\lambda f$ are admissible scores: their coordinate growth rates are strictly below the corresponding Gamma rates, so their normalizers and tilted first moments are finite. Integrating (40) from η to $\eta+\lambda$ as in Theorem 2 gives

$$\begin{aligned} \log \mathbb{E}_{Q_g} e^{\lambda(f-\mathbb{E}_P f)} &\le \sum_i \alpha_i \left[-\log\left(1-\frac{\lambda b_i}{1-\eta b_i}\right) - \lambda b_i\right] \\ &\le \lambda B_\eta + \frac{\lambda^2 V_\eta}{2(1-\lambda b_\eta)}. \end{aligned} \tag{43}$$

To verify the linear term, put $\widetilde{b}_i = b_i/(1-\eta b_i)$ and use $\sum_i \alpha_i(\widetilde{b}_i - b_i) = B_\eta$; the remaining terms have the ordinary sub-gamma form with b_i replaced by $\widetilde{b}_i$. Consequently,

$$Q_g\{f - \mathbb{E}_P f \ge B_\eta + t\} \le \exp\left(-\frac{t^2}{2(V_\eta + b_\eta t)}\right), \qquad t > 0.$$

The lower-tail bound follows by replacing f with $-f$, again giving a factor of two for the two-sided event. At $\eta = 0$ the parameters reduce to $B_0 = 0$, $V_0 = V$, and $b_0 = b$. For positive η, the displayed posterior bound explicitly accounts for the change in fluctuation scale as well as the population-centered bias.

I.i.d. averages and attainment. For n i.i.d. $\mathrm{Gamma}(\alpha,\beta)$ coordinates and a $1/n$-Lipschitz statistic f under ℓ_1, put $F = nf$. Apply (43) to the 1-Lipschitz function F at parameter λ/n, with $a_i = 1$. The product parameters B_η, V_η, b_η are thereby divided by n, n^2, n, respectively. In the rest of this paragraph these symbols denote the resulting sample-average parameters in (25). This gives (24). A transcript log-likelihood with sensitivity $\eta < \beta$ is an admissible initial score, so the calculation holds conditionally on almost every transcript. Relative to $\eta = 0$, the variance and large-deviation parameters grow by $(1-\eta/\beta)^{-2}$ and $(1-\eta/\beta)^{-1}$.

For the admissible score $g(x) = \eta \sum_i x_i$ and statistic $f(x) = n^{-1}\sum_i x_i$, the tilted coordinates are independent with law $\mathrm{Gamma}(\alpha, \beta-\eta)$. Direct calculation gives

$$\mathbb{E}_{Q_g} f - \mathbb{E}_P f = \frac{\alpha\eta}{\beta(\beta-\eta)} = B_\eta, \qquad \mathrm{Var}_{Q_g}(f) = \frac{\alpha}{n(\beta-\eta)^2} = V_\eta.$$

The first MGF bound in (24) is exact for this tilt. Thus the mean shift and variance increase are attained within the admissible tilt class; no claim is made that a specified selection mechanism realizes this score, or that the final Bernstein relaxation is exact.

E THE EXPONENTIAL-ENTROPY ROUTE AND ITS NORMALIZATION

For a positive random variable H, write

$$\mathrm{Ent}_P(H) = \mathbb{E}_P\left[H \log \frac{H}{\mathbb{E}_P H}\right].$$

Let Γ measure squared local sensitivity and put $\|u\|_\Gamma = \|\sqrt{\Gamma(u)}\|_\infty$. For example, $\Gamma(u) = \|\nabla u\|_2^2$ in a smooth Euclidean setting. Assume this seminorm is homogeneous and subadditive. Suppose that, for some $\rho > 0$ and a specified class $\mathcal{U}$,

$$\mathrm{Ent}_P(e^u) \le \frac{1}{2\rho}\mathbb{E}_P[e^u \Gamma(u)], \qquad u \in \mathcal{U}. \tag{44}$$

This is the exponential form of a logarithmic Sobolev inequality in smooth settings. Suitable modified log-Sobolev inequalities give analogous bounds in discrete settings, with constants determined by the convention for Γ.

Proposition 6 (Drift from exponential entropy)**.** *Suppose* $\|g\|_\Gamma \le \eta$, $\|f\|_\Gamma \le L$, *and* (44) *holds for* $u = \pm sf$ *and* $u = g \pm sf$ *for every* $s \ge 0$. *Assume* $f, g \in L^1(P)$, *these scores have finite positive exponential moments, and their tilts integrate* $|f| + |g|$. *Then, for* $\lambda \ge 0$,

$$|\mathbb{E}_{Q_g} f - \mathbb{E}_P f| \le \frac{L\eta}{\rho}, \qquad \log \mathbb{E}_{Q_g} e^{\lambda(f - \mathbb{E}_P f)} \le \frac{L\eta}{\rho}\lambda + \frac{L^2}{2\rho}\lambda^2. \tag{45}$$

If these hypotheses hold for every admissible Lipschitz score and every bounded Lipschitz test function, and $\|h\|_\Gamma \le \mathrm{Lip}_d(h)$, *then* $\Theta_{P,d}(\eta) \le \eta/\rho$.

We prove Proposition 6 using concentration under the original law and the entropy of the tilt. The stated path and integrability assumptions justify the identities and differentiation below.

Base concentration. Let $\Lambda(s) = \log \mathbb{E}_P e^{s(f - \mathbb{E}_P f)}$. The entropy identity and (44) at sf imply

$$s\Lambda'(s) - \Lambda(s) = \frac{\mathrm{Ent}_P(e^{sf})}{\mathbb{E}_P e^{sf}} \le \frac{s^2 L^2}{2\rho}.$$

For $s > 0$, divide by s^2 and integrate the derivative of $\Lambda(s)/s$. Its limit at zero is zero, because the statistic is centered. Thus

$$\log \mathbb{E}_P e^{s(f - \mathbb{E}_P f)} \le \frac{s^2 L^2}{2\rho}. \tag{46}$$

Applying the same argument to $-f$ gives the other sign.

Expectation drift. At the posterior score g, the entropy inequality gives

$$D_{\mathrm{KL}}(Q_g \| P) = \frac{\mathrm{Ent}_P(e^g)}{\mathbb{E}_P e^g} \le \frac{1}{2\rho}\mathbb{E}_{Q_g}\Gamma(g) \le \frac{\eta^2}{2\rho}. \tag{47}$$

The entropy variational inequality and (46) yield, for every $s > 0$,

$$\mathbb{E}_{Q_g} f - \mathbb{E}_P f \le \frac{D_{\mathrm{KL}}(Q_g \| P) + \log \mathbb{E}_P e^{s(f - \mathbb{E}_P f)}}{s} \le \frac{\eta^2}{2\rho s} + \frac{sL^2}{2\rho}.$$

For $\eta, L > 0$, take $s = \eta/L$ to obtain $L\eta/\rho$. The zero cases follow by taking a limit in s. Use $-f$ for the absolute-value bound.

Posterior concentration. The score $g + uf$ has seminorm at most $\eta + uL$. The preceding expectation bound, now applied to this score, gives

$$\frac{d}{du} \log \mathbb{E}_{Q_g} e^{u(f - \mathbb{E}_P f)} = \mathbb{E}_{Q_{g+uf}} f - \mathbb{E}_P f \le \frac{L(\eta + uL)}{\rho}.$$

Integration gives (45). To bound the full posterior drift profile, assume the path inequality applies to the requisite Lipschitz scores and test functions and that $\|h\|_\Gamma \le \mathrm{Lip}_d(h)$. First take bounded unit-Lipschitz test functions. Their base exponential moments are finite, and bounded perturbations of an admissible Lipschitz score retain the required integrability. The finite first moments of P and Q_g allow passage to arbitrary Lipschitz tests by truncation. Taking the suprema over these tests and admissible η-Lipschitz scores gives $\Theta_{P,d}(\eta) \le \eta/\rho$. If the functional inequality holds only on a restricted class, the same proof applies only to score–test pairs for which both families of paths are admitted.

Gaussian and strongly log-concave laws. For $P = N(0, \sigma^2 I)$, the Gaussian logarithmic Sobolev inequality (Gross, 1975) in exponential form is

$$\mathrm{Ent}_P(e^u) \le \frac{\sigma^2}{2}\mathbb{E}_P[e^u \|\nabla u\|_2^2].$$

Thus $\rho = 1/\sigma^2$ and $\Gamma(u) = \|\nabla u\|_2^2$. Lipschitz scores have the needed exponential integrability under the Gaussian law; approximation extends the inequality from smooth functions. Proposition 6 yields $\Theta(\eta) \le \sigma^2\eta$. For a unit vector v, the tilt with score $g(x) = \eta\langle v, x\rangle$ is $N(\sigma^2\eta v, \sigma^2 I)$.

Translation gives an upper coupling cost $\sigma^2\eta$, and the test function $\langle v, x\rangle$ gives the matching lower cost. This proves (22).

The logarithmic Sobolev inequality for a smooth density proportional to e^{-V} with $\nabla^2 V \succeq \kappa I$ has the same exponential form with coefficient $1/(2\kappa)$ (Bakry & Émery, 1985). Applying the same argument gives $\Theta(\eta) \le \eta/\kappa$. A stronger displacement bound is known: Khudiakova et al. (2025, Proposition 1.1) prove $W_\infty(Q_g, P) \le \mathrm{Lip}(g)/\kappa$ for strongly log-concave reference densities. Our argument records the entropy route to the W_1 profile. The factor two in the general T_1 comparison is unavoidable without further assumptions, as shown in Appendix F.

Reversible-chain conventions. Let K be a reversible Markov kernel on a finite or countable state space, with stationary law P, and define

$$\Gamma(u)(x) = \frac{1}{2}\sum_y K(x,y)(u(y)-u(x))^2,$$

$$\mathcal{E}(v,w) = \frac{1}{2}\sum_{x,y} P(x)K(x,y)(v(y)-v(x))(w(y)-w(x)).$$

Suppose the modified log-Sobolev convention is

$$\mathrm{Ent}_P(h) \le \rho_{\mathrm{ch}}^{-1}\mathcal{E}(h, \log h).$$

For real a, b, the elementary inequality

$$(e^a - e^b)(a-b) \le \tfrac{1}{2}(e^a + e^b)(a-b)^2$$

and reversibility imply

$$\begin{aligned}\mathcal{E}(e^u, u) &\le \frac{1}{4}\sum_{x,y} P(x)K(x,y)(e^{u(x)} + e^{u(y)})(u(y)-u(x))^2 \\ &= \mathbb{E}_P[e^u \Gamma(u)].\end{aligned}$$

Consequently (44) holds with $\rho = \rho_{\mathrm{ch}}/2$ under this convention. A restricted chain inequality yields this implication only for the corresponding admissible $h = e^u$. This normalization must be checked before importing an MLSI constant from another source.

F Transport comparison and sharpness

Proof of Proposition 5. For the forward implication, assume $\Theta_{P,d}(\eta) \le C\eta$ for all $\eta \ge 0$. Taking $g = 0$ in Theorem 2 and integrating this linear bound gives, for every 1-Lipschitz f,

$$\log \mathbb{E}_P e^{\lambda(f - \mathbb{E}_P f)} \le \frac{C\lambda^2}{2}, \qquad \lambda \in \mathbb{R}, \tag{48}$$

where negative λ follow by applying the theorem to $-f$. The log-Laplace characterization of T_1 (Bobkov & Götze, 1999) then yields $T_1(C)$. For the converse, fix an admissible score with $\mathrm{Lip}_d(g) \le \eta$ and put $w = W_{1,d}(Q_g, P)$. Since g is Lipschitz and both laws have finite first moments, $g \in L^1(P) \cap L^1(Q_g)$. Both relative entropies in the following identity are therefore finite:

$$D_{\mathrm{KL}}(Q_g \| P) + D_{\mathrm{KL}}(P \| Q_g) = \mathbb{E}_{Q_g} g - \mathbb{E}_P g \le \eta w. \tag{49}$$

Using $T_1(C)$ with reference law P gives

$$w^2 \le 2C D_{\mathrm{KL}}(Q_g \| P) \le 2C\big(\eta w - D_{\mathrm{KL}}(P \| Q_g)\big) \le 2C\eta w. \tag{50}$$

Divide by w when $w > 0$; the case $w = 0$ is immediate. Taking the supremum over admissible scores proves the bound. Sharpness is proved below. □

Why the converse loses a factor of two. The assumption controls transport using entropy relative to P. It does not provide the corresponding inequality with reference law Q_g. Below, a Bernoulli example rules out coefficient one in (54), and a family of such laws establishes the sharp factor two. In particular, the sub-Gaussian bound $\Lambda_f(\lambda) \le C\lambda^2/2$ cannot in general be differentiated to conclude $\Lambda_f'(\lambda) \le C\lambda$.

If Q_g also satisfies $T_1(C)$, the bound for this particular tilt improves to $w \le C\eta$. Applying transport in both directions and adding gives

$$w^2 \le C[D_{\mathrm{KL}}(Q_g\|P) + D_{\mathrm{KL}}(P\|Q_g)] \le C\eta w.$$

This improvement uses an additional assumption on the posterior.

All distributions in the remainder of this section are on $\{0,1\}$ with $d(0,1) = 1$.

The optimal Bernoulli transport constant. Let $P_p = \mathrm{Bernoulli}(p)$ with $0 < p < 1/2$, and define

$$a = \log\frac{1-p}{p}, \qquad m = 1-2p, \qquad C_p = \frac{m}{2a}. \tag{51}$$

This is the optimal Bernoulli sub-Gaussian constant in the Kearns–Saul inequality (Berend & Kontorovich, 2013), expressed as a T_1 constant through Bobkov & Götze (1999). We include a direct entropy proof that P_p satisfies $T_1(C_p)$ and that C_p is optimal. Every competing law is $R_q = \mathrm{Bernoulli}(q)$, and

$$W_1(R_q, P_p) = |q-p|, \qquad D_p(q) := D_{\mathrm{KL}}(R_q\|P_p) = q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}.$$

Thus it suffices to prove $F(q) := D_p(q) - (a/m)(q-p)^2 \ge 0$. Direct substitution gives $F(p) = F(1-p) = 0$. With $x = 2q-1$ and $a = 2\,\mathrm{artanh}(m)$, differentiation gives

$$F'(q) = 2\left[\mathrm{artanh}(x) - \frac{x}{m}\,\mathrm{artanh}(m)\right], \qquad 0 < q < 1. \tag{52}$$

The function

$$\frac{\mathrm{artanh}(x)}{x} = \int_0^1 \frac{dt}{1-x^2t^2}$$

is even and strictly increasing in $|x| \in [0,1)$, with value one at zero. Consequently, the signs of F' are respectively $-,+,-,+$ on the intervals separated by $p, 1/2, 1-p$. The two global minima are at p and $1-p$, both with value zero. Continuity includes the endpoints $q = 0, 1$. Therefore

$$W_1(R_q, P_p)^2 = (q-p)^2 \le \frac{m}{a} D_p(q) = 2C_p D_{\mathrm{KL}}(R_q\|P_p). \tag{53}$$

Equality at $q = 1-p \ne p$ proves optimality of C_p.

A counterexample to coefficient one. For $p = 1/10$, (51) gives $C_p = 1/(5\log 3)$. A linear score $g(x) = \eta x$ changes the Bernoulli parameter to

$$q_\eta = \frac{pe^\eta}{1-p+pe^\eta}.$$

At $\eta = \log 27$, this is $q_\eta = 3/4$, so

$$\Theta_{P_p,d}(\log 27) \ge W_1(Q_g, P_p) = \frac{3}{4} - \frac{1}{10} = \frac{13}{20} > \frac{3}{5} = C_p\eta. \tag{54}$$

Thus even a two-point reference measure with its optimal transport constant need not obey $\Theta_{P_p,d}(\eta) \le C_p\eta$.

The universal factor two is sharp. Let $a \to \infty$ and set

$$p_a = \frac{1}{1+e^a}, \qquad C_a = \frac{1-2p_a}{2a}, \qquad \eta_a = a + \sqrt{a}.$$

The reference law $P_a = \mathrm{Bernoulli}(p_a)$ satisfies $T_1(C_a)$ by (53). Tilting by $g_a(x) = \eta_a x$ gives a Bernoulli parameter

$$q_a = \frac{1}{1 + e^{-\sqrt{a}}}.$$

It follows that $W_1(Q_{g_a}, P_a) = q_a - p_a \to 1$, while

$$C_a \eta_a = \frac{1 - 2p_a}{2}\left(1 + \frac{1}{\sqrt{a}}\right) \longrightarrow \frac{1}{2}.$$

Together with the upper bound already proved in Proposition 5, this yields

$$2 \geq \frac{\Theta_{P_a,d}(\eta_a)}{C_a \eta_a} \geq \frac{q_a - p_a}{C_a \eta_a} \longrightarrow 2.$$

No smaller universal coefficient can therefore replace two in the converse implication.

G Comparison with prior work

Adaptive data analysis: separating the sampling model from the transfer proof. Theorem 1 makes the bias term $L\Theta_{P,d}(\eta)$ explicit: P and d describe the sampling model, L describes the statistic, and η describes the transcript likelihood. A change in the sampling model therefore calls for a new estimate of Θ, while the conditional-expectation transfer proof remains unchanged. We build here on the privacy-to-generalization connection of Dwork et al. (2015); Bassily et al. (2016) and the posterior decomposition of Jung et al. (2020). The latter already avoids the monitor argument and gives conditional concentration and simultaneous answer guarantees. The bounded-query estimate in (19) can also be recovered from their martingale argument (Appendix D, Lemma 22) by using conditional range lengths and the sharp one-coordinate total-variation bound. This example checks that the drift method retains the usual sampling scale. The contribution is the distributional profile, its exact product rule, and its use for both transfer and concentration across the models treated here. Appendix H handles their approximate-DP setting through the pure profile and an expected excess drift.

Metric privacy: an established stability model. Differential privacy originates in Dwork et al. (2006); Chatzikokolakis et al. (2013) formulate its extension to arbitrary metrics and study its Bayesian interpretation. Moreover, Shenfeld & Ligett (2023, Supplementary Appendix F.3) already interpret their adaptive-analysis method through metric privacy with query-dependent distances. We therefore do not claim to introduce non-Hamming geometry to adaptive data analysis. Our contribution in this setting is to express the transfer allowance through a profile of a fixed reference law and metric, calculate that profile, and reuse it for concentration. The metric-privacy assumption supplies the likelihood bound; tensorization and tilt integration supply the additional structure.

Posterior Wasserstein bounds and likelihood sensitivity. A direct precursor is the backward-channel approach to Wasserstein generalization bounds, developed by Lopez & Jog (2018) and discussed in Rodríguez-Gálvez et al. (2021, Section 3.3). For W_1, this approach bounds expected generalization by averaging the distances $W_1(P_{S|W=w}, P_S)$ between the data law conditional on a learned output and the original data law. The underlying Wasserstein duality already bounds the posterior mean shift separately for each output. Our profile $\Theta_{P,d}(\eta)$ takes a further supremum: it bounds this distance over all admissible tilts whose log-likelihood has Lipschitz constant at most η. This makes it a property of the reference law and metric, independent of a particular learner. We develop two structural results for this quantity. Exact tensorization computes the profile from coordinate profiles for product reference laws with additive metrics, even when the likelihood couples the coordinates (Theorem 3). Integrating the profile along further tilts $g + \lambda f$ yields conditional concentration around the original population mean (Theorem 2); a bound on the initial posterior distance alone does not control these additional tilts.

Related posterior stability notions include local statistical stability, which measures the total-variation change of a uniformly sampled record (Shenfeld & Ligett, 2019), and Bayes stability, which controls the selected query's posterior–prior expectation difference and supports variance-sensitive guarantees (Shenfeld & Ligett, 2023). The latter's posterior formulation also extends to

arbitrary dataset queries (Supplementary Appendix C.2, Lemma C.5). Our profile supplies a sufficient bound uniform over Lipschitz observables and likelihoods, together with the tensorization and concentration results above. The tradeoff is that the likelihood-sensitivity assumption restricts the learners covered, and taking a supremum can discard information specific to a learner or query. The benefit is therefore a way to calculate a distributional bound and use it for both transfer and concentration; it does not guarantee smaller bounds than an analysis of a particular posterior.

Information-theoretic and PAC-Bayesian generalization. Mutual information controls adaptive selection bias and generalization (Russo & Zou, 2016; Xu & Raginsky, 2017), while conditional mutual information connects several approaches to learning guarantees (Steinke & Zakynthinou, 2020). These methods are not limited to expectations: information-density bounds also give high-probability, single-draw, and PAC-Bayesian guarantees (Hellström & Durisi, 2020). Classical PAC-Bayesian bounds concern data-dependent distributions over hypotheses (McAllester, 1999); our posterior is instead the conditional law of the dataset given an output. Our pointwise likelihood-sensitivity assumption is restrictive, but lets us use a reference-law profile to obtain concentration conditional on almost every output. This is a different assumption and mode of control, rather than a uniform improvement over information-theoretic guarantees.

Dependent and unbounded data: explicit error bounds. For a uniform K-subset of $[N]$, consider a transcript whose log-likelihood is η-Lipschitz under swap distance and which selects a $[0,1]$-valued function to average over the sample. Proposition 3 gives selection bias at most $(m/K)\tanh(\eta/4)$ and a sampling term $\sqrt{m\log(2/\delta)/(2K^2)}$, where $m = \min(K, N-K)$. Thus the same transfer and concentration theorems give an explicit guarantee despite dependence among the sampled elements. For unbounded Gamma observations with rate β and likelihood sensitivity $0 \le \eta < \beta$, (25) additionally shows how selection changes the tail: the variance parameter increases by $(1-\eta/\beta)^{-2}$ and the large-deviation scale by $(1-\eta/\beta)^{-1}$, relative to the unselected bound. Prior work also handles dependent data. Typical stability takes fixed-query concentration as an input to adaptive generalization (Bassily & Freund, 2016), whereas our drift calculation establishes both fixed-query concentration and its conditional version. Kontorovich et al. (2022) use a Gibbs-dependence penalty in their transfer guarantee; our slice calculation instead works directly with the sampling law and swap distance. The advantage is a direct route from a model-specific expectation bound to both conclusions, with the cost of selection visible in the resulting parameters. It does not require a separate fixed-query concentration proof, but it does require estimating the profile for the model in question.

Stability-based concentration: removing the auxiliary selection step. Relative to the privacy-based concentration proof of Steinke & Ullman (2017), our method removes an auxiliary optimization problem. Their argument constructs a stable approximate maximizer over independent repetitions and controls the expected selected maximum. Once a drift bound is available, Theorem 2 instead differentiates the log-moment generating function, bounds the resulting expectation shift, and integrates. For bounded statistics this proof is a single derivative calculation; truncation extends it to unbounded Lipschitz statistics. It recovers the usual bounded-differences constant through Proposition 2 and also handles the finite exponential-moment range of Gamma data. Starting from a transcript posterior only changes the likelihood budget along the integration path, from uL to $\eta+uL$. Thus the ordinary and adaptive bounds use the same calculation. Nissim & Stemmer (2019) obtain a different scope advantage by allowing high worst-case sensitivity under conditions on the tails of coordinate changes; our global Lipschitz assumption does not cover all of those cases.

Local geometry: a drift formulation of Talagrand-type concentration. State-dependent coordinate weights and one-sided transport costs are established tools in concentration. Talagrand's convex-distance theory yields dimension-free product-space inequalities (Talagrand, 1995); Marton's asymmetric transport approach and subsequent weak transport theory extend such tools to dependent data and more general costs (Marton, 1996; Wintenberger, 2015). Section 6 uses this geometry to define a one-sided local posterior drift profile. Its specific input is a local certificate for the log-likelihood, and its output is the expectation comparison $\mathbb{E}_{Q_g} f - \mathbb{E}_P f \le L\eta$. Closure of the certificate classes under $g + \lambda f$ then yields adaptive upper-tail concentration around the original mean, with bias allowance $L\eta$ and variance proxy L^2. This is the advantage within our framework: the same drift-and-integration argument handles both fixed metric sensitivity and local, asymmetric

sensitivity. The norm example shows an improvement over global bounded differences, not over the existing theory of convex concentration. We do not claim the sharp classical convex-distance set inequality or a new exact tensorization identity for the local profile. The coupling proof also recovers the upper half of our metric tensorization theorem; the matching product-tilt construction is still needed for equality.

Direct bounds for likelihood perturbations. Wasserstein comparisons under changes of density are also established. In one dimension, Ley et al. (2017) bound the distance between nested densities using the derivative of their density ratio and a Stein kernel; this gives Gaussian drift bounds and an alternative route to the Gamma upper bound under their regularity and boundary conditions. For a κ-strongly log-concave reference density, Khudiakova et al. (2025, Proposition 1.1) prove $W_\infty(Q_g, P) \le \mathrm{Lip}(g)/\kappa$, which is stronger than the W_1 bound used here and is sharp for Gaussian translations. Thus these model-specific displacement bounds are not new consequences unique to our method. Our framework places them in a common profile calculus that also covers discrete and dependent examples and connects them to adaptive transfer.

Entropy and transport: conditional bounds from the original law. Once a drift bound for the original law is proved, we obtain a conditional tail bound without proving a new log-Sobolev or transportation inequality for the transcript posterior. For example, for Gaussian reference data with covariance $\sigma^2 I$, the exact profile $\Theta(\eta) = \sigma^2\eta$ gives a conditional radius

$$\underbrace{\sigma^2 L\eta}_{\text{selection-bias allowance}} + \underbrace{\sigma L\sqrt{2\log(2/\delta)}}_{\text{sampling term}}$$

around $\mathbb{E}_P f_\pi$, with confidence $1-\delta$ conditional on almost every transcript, for an L-Lipschitz statistic selected by a transcript whose log-likelihood is η-Lipschitz (Corollary 1). The posterior need not be Gaussian or satisfy the original log-Sobolev inequality.

The mechanism behind this conclusion differs from the usual Herbst step: the entropy method controls $\lambda\Lambda'(\lambda) - \Lambda(\lambda)$ for a log-moment generating function Λ (Boucheron et al., 2013), while we bound $\Lambda'(\lambda)$ directly by the original law's drift profile evaluated at $\eta + \lambda L$. Proposition 6 turns suitable entropy inequalities into this drift bound, including restricted inequalities when they hold along the required paths. Restricted concentration arguments appear in Anari et al. (2021, Section 7), with a merged conference version in Anari et al. (2022); their hypotheses must be checked for the posterior tilt paths used here. Transport maps onto log-Lipschitz perturbations can also transfer functional inequalities under curvature and regularity assumptions (Fathi et al., 2024); our population-centered tail bound requires no such map. Connections between expected generalization bounds and transportation-cost inequalities through convex duality were developed by Esposito & Gastpar (2022). Our comparison concerns the worst drift over log-Lipschitz tilts: a linear profile with coefficient C implies $T_1(C)$, whereas the reverse implication incurs a sharp factor of two (Proposition 5, using Bobkov & Götze, 1999). Keeping the full profile additionally handles the Gamma example, where no finite global linear bound exists. Classical entropy and transport methods also treat nonlinear tails; the contribution here is the explicit passage from one expectation-drift calculation to transfer, ordinary concentration, and concentration after selection.

H Approximate differential privacy through excess posterior drift

Approximate differential privacy does not give the pointwise likelihood constraint used in Section 3. Nevertheless, on product data the same pure-privacy profile remains useful: it bounds a baseline drift, and the privacy parameter δ controls the expected excess beyond that baseline. We prove this through a one-coordinate robustness bound and the product decomposition from Appendix A. The resulting transfer theorem uses the same conditional triangle inequality as Theorem 1.

Approximate-DP posterior transfer for statistical and low-sensitivity queries already appears in Jung et al. (2020, Lemma 7 and Appendix A.1). The purpose here is to express this extension in terms of our distribution-dependent profile and an explicit defect.

H.1 A ONE-COORDINATE ROBUSTNESS BOUND

For a probability law P_0 on a standard Borel space, write $\mathrm{TV}(\nu, P_0) = \sup_A |\nu(A) - P_0(A)|$ and define

$$\theta_{P_0}(\epsilon) := \sup_{\mathrm{osc}_{P_0}(g) \le \epsilon} \mathrm{TV}(Q_g, P_0), \qquad Q_g = \frac{e^g P_0}{\mathbb{E}_{P_0} e^g}, \tag{55}$$

Here $\mathrm{osc}_{P_0}(g)$ is the essential supremum of g minus its essential infimum under P_0; scores are measurable and real-valued. On a finite or countable space with discrete metric $d_{\mathrm{disc}}(x, y) = \mathbf{1}\{x \neq y\}$, this is exactly $\Theta_{P_0, d_{\mathrm{disc}}}(\epsilon)$, after restricting to the support. The total-variation formulation also applies on general standard Borel spaces without imposing a discrete topology. This coordinate profile has the elementary formula

$$\theta_{P_0}(\epsilon) = \sup_A \frac{(e^\epsilon - 1)P_0(A)(1 - P_0(A))}{1 + (e^\epsilon - 1)P_0(A)}. \tag{56}$$

Indeed, every admissible tilt satisfies $Q_g(A) \le e^\epsilon P_0(A)/[1 + (e^\epsilon - 1)P_0(A)]$, which gives the upper bound after taking the supremum over A. The score $g = \epsilon \mathbf{1}_A$ attains the corresponding set difference, giving the reverse bound. Maximizing over all set masses in $[0, 1]$ recovers the estimate from Appendix B:

$$\theta_{P_0}(\epsilon) \le b_\epsilon := \tanh(\epsilon/4) < 1 \qquad (0 \le \epsilon < \infty). \tag{57}$$

For probability measures U, V and $c \ge 1$, define the hockey-stick divergence

$$D_c(U \| V) := \sup_A \{U(A) - cV(A)\}.$$

For an arbitrary probability measure ν, let

$$h_\epsilon(\nu, P_0) := D_{e^\epsilon}(\nu \otimes P_0 \,\|\, P_0 \otimes \nu). \tag{58}$$

This quantity lies in $[0, 1]$ and measures a violation of the pure likelihood-ratio constraint. No absolute continuity of ν with respect to P_0 is required.

Lemma 1 (Drift beyond the pure profile)**.** *Let $\epsilon \ge 0$ be finite and choose any deterministic $\theta \in [\theta_{P_0}(\epsilon), 1)$. For every probability measure ν,*

$$\big(\mathrm{TV}(\nu, P_0) - \theta\big)_+ \le (1 - \theta) h_\epsilon(\nu, P_0). \tag{59}$$

Proof. Write $v = \mathrm{TV}(\nu, P_0)$ and $c = e^\epsilon$. Choose a positive Hahn set A for $\nu - P_0$, so that $v = \nu(A) - P_0(A)$, and put $p = P_0(A)$ and $k = 1 + (c - 1)p$. The pure tilt with score $g = \epsilon \mathbf{1}_A$ has drift

$$v_p := Q_g(A) - P_0(A) = \frac{(c - 1)p(1 - p)}{k} \le \theta_{P_0}(\epsilon) \le \theta.$$

Testing the divergence in (58) on $A \times A^c$ yields

$$h_\epsilon(\nu, P_0) \ge (p + v)(1 - p) - cp(1 - p - v) = k(v - v_p).$$

If $v \le \theta$, the claim is immediate. Otherwise $\theta < v \le 1 - p$. The affine function

$$F(t) = k(t - v_p) - \frac{t - \theta}{1 - \theta}$$

is nonnegative at both ends of $[\theta, 1 - p]$: at $t = \theta$ this follows from $v_p \le \theta$, and at the other endpoint

$$F(1 - p) = (1 - p) - \frac{1 - p - \theta}{1 - \theta} = \frac{p\theta}{1 - \theta} \ge 0.$$

Hence $k(v - v_p) \ge (v - \theta)/(1 - \theta)$, proving the result. □

The coefficient $1 - \theta$ is sharp uniformly over reference laws and posteriors: for nonatomic P_0 and a point mass ν, both $\mathrm{TV}(\nu, P_0)$ and $h_\epsilon(\nu, P_0)$ equal one.

H.2 Approximate privacy bounds the average defect

Consider a one-record channel K_x with a standard Borel output space. Suppose that, for all inputs x, y and measurable output events A,

$$K_x(A) \le e^{\epsilon} K_y(A) + \delta. \tag{60}$$

Let $X \sim P_0$, let $\Pi \sim K_X$, and write $\nu_\pi = \mathcal{L}(X \mid \Pi = \pi)$.

Lemma 2 (Average posterior defect)**.** *Under* (60)*,*

$$\mathbb{E}_\Pi h_\epsilon(\nu_\Pi, P_0) \le \delta. \tag{61}$$

Consequently, for every $\theta \in [\theta_{P_0}(\epsilon), 1)$*,*

$$\mathbb{E}_\Pi\big(\mathrm{TV}(\nu_\Pi, P_0) - \theta\big)_+ \le (1-\theta)\delta.$$

Proof. Compare the probability measures

$$\mathsf{R}(dx, dy, d\pi) = P_0(dx)P_0(dy)K_x(d\pi), \qquad \mathsf{S}(dx, dy, d\pi) = P_0(dx)P_0(dy)K_y(d\pi).$$

For any event in the joint space, apply (60) to its output section at each (x, y) and integrate. This gives $D_{e^\epsilon}(\mathsf{R}\|\mathsf{S}) \le \delta$. Both measures have the same transcript marginal, say $m(d\pi)$. Their conditional laws of (X, Y) are respectively $\nu_\pi \otimes P_0$ and $P_0 \otimes \nu_\pi$.

For kernels on standard Borel spaces, hockey-stick divergence disintegrates over a common marginal:

$$D_c\big(m(d\pi)U_\pi \,\|\, m(d\pi)V_\pi\big) = \int D_c(U_\pi\|V_\pi)\, m(d\pi).$$

This identity follows by taking the defining supremum over a countable generating algebra and choosing measurable approximate maximizers. Applying it proves (61); Lemma 1 gives the remaining assertion, including for singular posteriors. □

H.3 Product data and accuracy transfer

Let $X = (X_1, \dots, X_n) \sim P = \bigotimes_i P_i$, with standard Borel coordinate and transcript spaces. Let the full transcript channel K_x be (ϵ, δ)-DP under replacement of one coordinate, where $0 \le \epsilon < \infty$ and $0 \le \delta \le 1$. Write $Q_\pi = \mathcal{L}(X \mid \Pi = \pi)$ and choose deterministic bounds

$$\theta_i \in [\theta_{P_i}(\epsilon), 1), \qquad i = 1, \dots, n.$$

Selected queries $(x, \pi) \mapsto f_{j,\pi}(x)$ are real-valued and jointly measurable, and $j = 1, \dots, k$. Assume a common deterministic sensitivity envelope:

$$\mathrm{osc}_i(f_{j,\pi}) \le c_i < \infty, \qquad c_i \ge 0, \qquad B := \sum_i c_i\theta_i, \quad C := \sum_i c_i. \tag{62}$$

The oscillations are global: the bounds hold for every transcript and every pair of datasets differing only at coordinate i. This matters because a posterior can charge a transcript-dependent prior-null set.

Theorem 5 (Product drift with an approximate-privacy defect)**.** *Under these assumptions, there is a measurable* $R(\pi) \ge 0$*, independent of the choice of queries within the sensitivity envelope, such that*

$$\max_{j \le k} |\mathbb{E}_{Q_\pi} f_{j,\pi} - \mathbb{E}_P f_{j,\pi}| \le B + R(\pi) \quad \textit{for almost every } \pi, \qquad \mathbb{E}R(\Pi) \le \delta(C - B). \tag{63}$$

Proof. Fix a coordinate i and a prior prefix $z = x_{<i}$. Average the channel over the independent prior suffix:

$$K_{i,z}(A \mid u) := \int K_{(z,u,y_{>i})}(A)\, P_{>i}(dy_{>i}).$$

For every fixed suffix, the two inputs $(z, u, y_{>i})$ and $(z, v, y_{>i})$ differ in at most one coordinate. Integrating their DP inequality shows that $K_{i,z}$ satisfies (60) with the same (ϵ, δ). Let $m_{i,z}$ be its output law when $u \sim P_i$, and choose a jointly measurable posterior kernel

$$\nu_{i,z,\pi} = \mathcal{L}(X_i \mid X_{<i} = z, \Pi = \pi), \qquad v_i(z, \pi) = \mathrm{TV}(\nu_{i,z,\pi}, P_i).$$

Standard Borel spaces admit countable generating algebras, so v_i is measurable as a countable supremum of measurable set differences. Using the same countable classes, choose one full-measure set of transcripts on which these kernels are the conditional coordinate laws of Q_π, simultaneously for all i. This set does not depend on the selected queries. For $P_{<i}$-almost every z, Lemma 2 gives

$$\int (v_i(z,\pi)-\theta_i)_+ \, m_{i,z}(d\pi) \le (1-\theta_i)\delta. \tag{64}$$

These measurable kernels only need to be specified up to the corresponding joint null sets.

For each fixed transcript and query, use its Doob decomposition under the original product law:

$$F_{j,i,\pi}(x_{\le i}) := \int f_{j,\pi}(x_{\le i}, y_{>i})\, P_{>i}(dy_{>i}), \qquad F_{j,0,\pi} = \mathbb{E}_P f_{j,\pi}.$$

The increment $D_{j,i,\pi} = F_{j,i,\pi} - F_{j,i-1,\pi}$ has P_i-mean zero for each prefix and oscillation at most c_i as a function of x_i. Thus

$$\left| \int D_{j,i,\pi}(z,u)\, \nu_{i,z,\pi}(du) \right| \le c_i v_i(z,\pi).$$

Summing the increments and averaging under the posterior prefix laws gives, simultaneously for all selected queries,

$$|\mathbb{E}_{Q_\pi} f_{j,\pi} - \mathbb{E}_P f_{j,\pi}| \le \sum_i c_i \mathbb{E}[v_i(X_{<i},\pi) \mid \Pi = \pi] \le B + R(\pi),$$

where

$$R(\pi) := \sum_i c_i \mathbb{E}[(v_i(X_{<i},\pi)-\theta_i)_+ \mid \Pi = \pi].$$

Finite coordinate oscillations ensure that each fixed query has finite total oscillation, so these integrals are well defined. Finally, averaging a posterior-prefix integral over Π restores the joint law $P_{<i}(dz)m_{i,z}(d\pi)$. Consequently,

$$\begin{aligned}\mathbb{E}R(\Pi) &= \sum_i c_i \int P_{<i}(dz) \int (v_i(z,\pi)-\theta_i)_+ \, m_{i,z}(d\pi) \\ &\le \delta \sum_i c_i(1-\theta_i) = \delta(C-B),\end{aligned}$$

by (64). □

The existing Wasserstein profile appears exactly. Suppose the coordinate spaces are finite or countable and take $\theta_i = \Theta_{P_i,d_{\mathrm{disc}}}(\epsilon)$. By Theorem 3,

$$T := \Theta_{P,d_H}(\epsilon) = \sum_i \theta_i, \qquad d_H(x,y) = \sum_i \mathbf{1}\{x_i \ne y_i\}.$$

Countability ensures that the Hamming space is Polish, as required by our metric formulation. For each transcript in the common full-measure set fixed in the proof above, the preceding decomposition with $c_i = 1$ bounds the drift of every 1-Lipschitz function by the same $T + R(\pi)$. Taking the supremum and using Wasserstein duality gives

$$\mathbb{E}_\Pi\big(W_{1,d_H}(Q_\Pi, P) - \Theta_{P,d_H}(\epsilon)\big)_+ \le \delta\big(n - \Theta_{P,d_H}(\epsilon)\big). \tag{65}$$

Thus the original profile supplies the baseline, with no new approximate profile needed. More generally, for queries with a common Hamming Lipschitz constant L, use $c_i = L$ to obtain $B = LT$ and $C = nL$. Unequal c_i are handled by (62). When all $c_i > 0$, their weighted sum should not be confused with $\Theta_{P,d_c}(\epsilon)$ for $d_c = \sum_i c_i \mathbf{1}\{x_i \ne y_i\}$, which would impose different coordinate likelihood budgets.

Corollary 4 (Approximate-DP accuracy transfer)**.** *Under the assumptions of Theorem 5, suppose the selected queries and measurable answers $a_j(\pi)$ take values in $[0,1]$ and, for $\alpha, \beta \in [0,1]$,*

$$\mathbb{P}\left[\max_{j\le k} |a_j(\Pi) - f_{j,\Pi}(X)| > \alpha\right] \le \beta.$$

Then, for every $u > 0$,

$$\mathbb{P}_\Pi\left[\max_{j\le k} |a_j(\Pi) - \mathbb{E}_P f_{j,\Pi}| > \alpha + B + u\right] \le \frac{(1-\alpha)\beta + \delta(C-B)}{u}. \tag{66}$$

Proof. Let $q(\pi)$ be the conditional probability of the joint sample-failure event. Then $\mathbb{E}q(\Pi) \le \beta$, and the conditional triangle inequality used in Theorem 1 gives

$$\max_j |a_j(\pi) - \mathbb{E}_P f_{j,\pi}| \le \alpha + (1-\alpha)q(\pi) + B + R(\pi).$$

Apply Markov's inequality to the nonnegative quantity $(1-\alpha)q(\Pi) + R(\Pi)$ and use (63). □

There is no query-count factor beyond that in the assumed joint sample accuracy: the same residual controls every query in the sensitivity envelope. At $\delta = 0$ it vanishes almost surely, recovering pure-profile transfer.

H.4 Statistical queries and the scope of the extension

For $X \sim P_0^{\otimes n}$ and statistical queries $f_{j,\pi}(X) = n^{-1}\sum_i \phi_{j,\pi}(X_i)$ with $0 \le \phi_{j,\pi} \le 1$, take $c_i = 1/n$. Then $C = 1$ and $B = \theta_{P_0}(\epsilon)$ when the exact coordinate profile is used. In particular, its value supplies the same bias allowance as in the pure-DP calculation. Choosing instead the universal bound $\theta_i = b_\epsilon$ in Corollary 4 yields

$$\begin{aligned}\mathbb{P}_\Pi\left[\max_{j\le k} |a_j(\Pi) - \mathbb{E}_{P_0}\phi_{j,\Pi}| > \alpha + b_\epsilon + u\right] \\ \le \frac{(1-\alpha)\beta + (1-b_\epsilon)\delta}{u} \le \frac{\beta+\delta}{u}, \qquad b_\epsilon = \tanh(\epsilon/4).\end{aligned} \tag{67}$$

Why the correction is an average defect. Approximate privacy cannot generally be reduced to the pure likelihood condition by discarding a small set of transcripts. For a one-record domain $[m]$, $m \ge 2$, let K_x be uniform on $[m] \setminus \{x\}$. For $x \ne y$, $\mathrm{TV}(K_x, K_y) = 1/(m-1)$, so the channel is $(0, 1/(m-1))$-DP. Under a prior with full support, every transcript t has zero likelihood at $x = t$ and positive likelihood elsewhere. Thus every transcript fails every finite global log-likelihood oscillation bound. Lemma 1 handles such posteriors by controlling their excess drift directly.

Limits of the present extension. The extension covers nonlinear dataset queries and independent coordinates with different laws, under deterministic global coordinate sensitivity bounds. Unbounded transportation costs need further distance or moment control: a probability-δ disclosure can carry arbitrarily large cost. The expected defect suffices for accuracy transfer; conditional concentration additionally needs control along the further exponential tilts in Section 4.

I Proofs and examples for one-sided local profiles

I.1 Proof of the local drift bound

We prove Theorem 4. Let $P = \bigotimes_{i=1}^n P_i$ on a finite product of standard Borel spaces, and let h, g have measurable nonnegative certificates a, b satisfying

$$h(x) - h(y) \le \sum_i a_i(x)\mathbf{1}_{\{x_i \ne y_i\}}, \qquad g(x) - g(y) \le \sum_i b_i(x)\mathbf{1}_{\{x_i \ne y_i\}},$$

with $\|a(x)\|_2 \le L$ and $\|b(x)\|_2 \le \eta$ for every x. These assumptions give $\mathrm{osc}(h) \le L\sqrt{n}$ and $\mathrm{osc}(g) \le \eta\sqrt{n}$, by Cauchy–Schwarz and interchange of x, y. Thus $Q = Q_g$ is well defined, its density $r = e^g/\mathbb{E}_P e^g$ is bounded above and away from zero, and all expectations below are finite.

Conditional likelihoods. For each prefix define

$$W_i(x_{\le i}) = \int r(x_{\le i}, z_{>i})\, P_{>i}(dz_{>i}), \qquad W_0 = 1.$$

The conditional law $Q_i(\cdot \mid x_{<i})$ has density

$$\rho_i(u \mid x_{<i}) = \frac{W_i(x_{<i}, u)}{W_{i-1}(x_{<i})} \quad \text{relative to } P_i.$$

Write $\bar{b}_i(X_{\le i}) = \mathbb{E}_Q[b_i(X) \mid X_{\le i}]$, using the version obtained by integrating $b_i r/W_i$ against $P_{>i}$. For every u, v and suffix $z_{>i}$, the certificate for g gives

$$r(x_{<i}, v, z_{>i}) \ge r(x_{<i}, u, z_{>i}) e^{-b_i(x_{<i}, u, z_{>i})}.$$

Integrating over the suffix and applying Jensen's inequality yields

$$\begin{aligned} W_i(x_{<i}, v) &\ge W_i(x_{<i}, u) \mathbb{E}_Q[e^{-b_i(X)} \mid X_{\le i} = (x_{<i}, u)] \\ &\ge W_i(x_{<i}, u) e^{-\bar{b}_i(x_{<i}, u)}. \end{aligned}$$

Integrating this inequality over $v \sim P_i$ proves

$$\rho_i(u \mid x_{<i}) \le e^{\bar{b}_i(x_{<i}, u)}. \tag{68}$$

A sequential coupling. Given the previously generated $(X_{<i}, Y_{<i})$, maximally couple $Q_i(\cdot \mid X_{<i})$ and P_i to generate (X_i, Y_i), using a kernel that depends only on $X_{<i}$. Such measurable kernels follow from the common density $\min\{\rho_i, 1\}$ relative to P_i and the two residual measures. The first marginal gives $X \sim Q$. The conditional law of Y_i given the full preceding history is P_i, so $Y \sim P$. For this maximal coupling,

$$\Pr(Y_i \ne X_i \mid X_{\le i}) = \left(1 - \rho_i(X_i \mid X_{<i})^{-1}\right)_+.$$

Moreover, the conditional marginal of each future X_j, given the preceding pairs, is $Q_j(\cdot \mid X_{<j})$ and does not depend on the Y history. Iterating these marginals shows that Y_i and $X_{>i}$ are conditionally independent given $X_{\le i}$. Consequently, conditioning on the full X preserves the displayed mismatch probability. By (68),

$$p_i(X) := \Pr(Y_i \ne X_i \mid X) \le 1 - e^{-\bar{b}_i(X_{\le i})} \le \bar{b}_i(X_{\le i}). \tag{69}$$

From the coupling to drift. Applying the certificate for h to the coupled pair gives

$$\mathbb{E}_Q h - \mathbb{E}_P h \le \mathbb{E} \sum_i a_i(X) \mathbf{1}_{\{X_i \ne Y_i\}} \le \mathbb{E}_Q \sum_i a_i(X) \bar{b}_i(X_{\le i}).$$

Let $\bar{a}_i(X_{\le i}) = \mathbb{E}_Q[a_i(X) \mid X_{\le i}]$. Conditional expectation followed by Cauchy–Schwarz gives

$$\mathbb{E}_Q \sum_i a_i \bar{b}_i = \mathbb{E}_Q \sum_i \bar{a}_i \bar{b}_i \le \left(\mathbb{E}_Q \sum_i \bar{a}_i^2\right)^{1/2} \left(\mathbb{E}_Q \sum_i \bar{b}_i^2\right)^{1/2}.$$

Conditional Jensen's inequality implies

$$\mathbb{E}_Q \sum_i \bar{a}_i^2 \le \mathbb{E}_Q \sum_i a_i^2 \le L^2, \qquad \mathbb{E}_Q \sum_i \bar{b}_i^2 \le \mathbb{E}_Q \sum_i b_i^2 \le \eta^2.$$

Hence $\mathbb{E}_Q h - \mathbb{E}_P h \le L\eta$. Taking $L = 1$ and the suprema in the definition of the local profile proves $\Theta_P^{\mathrm{loc},+}(\eta) \le \eta$. □

The same coupling also records a directed transport estimate. If $\Gamma(Q, P)$ denotes the couplings with $X \sim Q$ and $Y \sim P$, then

$$\inf_{\pi \in \Gamma(Q,P)} \mathbb{E}_Q \sum_i \Pr_\pi(Y_i \ne X_i \mid X)^2 \le \mathbb{E}_Q \sum_i \bar{b}_i^2 \le \eta^2.$$

The conditioning on X makes this quantity directional and retains the dependence of the certificate weights on the realized data. Appendix A.4 gives the corresponding metric construction, using coordinate transport costs to prove the upper tensorization bound.

I.2 From local drift to ordinary and adaptive concentration

The certificate in Definition 2 controls a decrease from x by weights evaluated at x. The cost $c_a(x, y) = \sum_i a_i(x) \mathbf{1}_{\{x_i \ne y_i\}}$ can be asymmetric and need not satisfy the triangle inequality. The statistic and score may use different certificate fields. The all-pairs requirement is essential: a bound

only for single-coordinate changes does not by itself imply (27) with the weights at the starting dataset.

The oscillation bound established above ensures finite exponential moments. Adding certificates and using the triangle inequality for their Euclidean norms gives

$$\mathcal{C}_\eta + u\mathcal{C}_L \subseteq \mathcal{C}_{\eta+uL}, \qquad u \geq 0.$$

This closure under positive tilts replaces the Lipschitz inequality $\mathrm{Lip}(g+uf) \leq \eta+uL$ in Section 4.

Proof of Corollary 3. For $g \in \mathcal{C}_\eta$ and $f \in \mathcal{C}_L$, the score $g + uf$ belongs to $\mathcal{C}_{\eta+uL}$. Theorem 4 gives

$$\frac{d}{du} \log \mathbb{E}_{Q_g} e^{u(f-\mathbb{E}_P f)} = \mathbb{E}_{Q_{g+uf}} f - \mathbb{E}_P f \leq L(\eta + uL).$$

Integration proves (30); Chernoff's bound with $\lambda = t/L^2$ proves (31) for $L > 0$. For $L = 0$, f is constant and its deviation is zero. □

Taking $g = 0$ gives ordinary upper-tail concentration. For adaptive use, let Π be a standard Borel transcript with jointly measurable likelihood densities against a common σ-finite measure. Suppose that, for almost every π, its log-likelihood g_π belongs to $\mathcal{C}_{\eta(\pi)}$ and its selected statistic f_π belongs to $\mathcal{C}_{L_\pi}$. The corollary then holds conditional on $\Pi = \pi$, with fresh-data mean $\mathbb{E}_P f_\pi$, bias allowance $L_\pi\eta(\pi)$, and fluctuation scale L_π. For jointly measurable selected functions and measurable certificates and bounds, choosing a conditional confidence level and averaging gives the corresponding unconditional guarantee. The likelihood condition must hold for the full transcript that determines the selected statistic. In general, a certificate for f_π does not give one for $-f_\pi$; lower tails require that separate certificate.

I.3 Examples and the connection to Talagrand's convex distance

A dimension-free fluctuation scale. On $[0,1]^n$, consider $f(x) = \|x\|_2$. For $x \neq 0$, the subgradient inequality gives

$$f(x) - f(y) \leq \frac{x}{\|x\|_2} \cdot (x - y) \leq \sum_i \frac{x_i}{\|x\|_2} \mathbf{1}_{\{x_i \neq y_i\}}.$$

Thus $a_i(x) = x_i/\|x\|_2$, with $a(0) = 0$, is a certificate of norm at most one. Each coordinate can have weight one somewhere, but the squared weights sum to at most one at any single dataset. Corollary 3 gives

$$Q_g[\|X\|_2 - \mathbb{E}_P\|X\|_2 \geq \eta + t] \leq e^{-t^2/2} \qquad (g \in \mathcal{C}_\eta).$$

The fluctuation scale is independent of n, whereas the global coordinate bounds $c_i = 1$ give $e^{-2t^2/n}$ at $g = 0$. The same certificate argument applies to $f_J(x) = \|x_J\|_2$, so a transcript may select the subset J when its log-likelihood has the required certificate. In the adaptive bound, the bias allowance η still depends on that likelihood certificate and may grow with n.

Talagrand's convex distance. This local geometry is classical in product-space concentration (Talagrand, 1995; Boucheron et al., 2013). Take finite coordinate spaces and a nonempty set $A \subseteq \Omega$. Talagrand's convex distance can be written as

$$d_T(x, A) = \max_{\alpha \in \mathbb{R}^n_+ : \|\alpha\|_2 \leq 1} \min_{z \in A} \sum_i \alpha_i \mathbf{1}_{\{x_i \neq z_i\}}.$$

A maximizing vector α_x supplies a certificate for $d_T(\cdot, A)$: the triangle inequality for the weighted Hamming cost gives

$$d_T(x, A) - d_T(y, A) \leq \sum_i \alpha_{x,i} \mathbf{1}_{\{x_i \neq y_i\}}.$$

Thus $d_T(\cdot, A) \in \mathcal{C}_1$, and the corollary controls its upper tail around $\mathbb{E}_P d_T(X, A)$, both under P and after a tilt with a local likelihood certificate. This is a centered functional bound; it does not establish the sharp classical set inequality with its $P(A)$ factor. The profile connects this established geometry to likelihood-controlled expectation drift and adaptive concentration. Appendix G compares this result with the weak transport literature.