Anytime-Valid Confirmation of Label-Shift Corrections

Seungjin Choi [1]

Abstract

In small-batch scientific deployments, labeled target outcomes may be too scarce for reliable shift estimation even when unlabeled target inputs are available. We address the complementary setting where the practitioner has a pre-specified label-shift correction from domain knowledge and asks whether incoming labeled outcomes support it. We show that the per-observation likelihood ratio between a label-shift-corrected predictive and the source predictive is a conditional e-value, so its running product is a nonnegative martingale and Ville's inequality yields an anytime-valid confirmation rule. The log martingale equals the cumulative negative log-predictive density (NLPD) gap between the source and the corrected predictive, converting routine model monitoring into a formal sequential test. Rejection means the incoming data support the posited correction relative to the source predictive, but it is not a precise estimate of the degree of shift. Closed forms are available for GP sources with Gaussian label-shift ratios. GP regression simulations validate Type I control, finite-sample power, miscalibration sensitivity, and the small-batch advantage of a reliable prior over label-based re-estimation.

1. Introduction

When deploying a predictive model in a new environment, the label distribution may shift between source and target domains, a phenomenon known as *label shift*, where the conditional distribution $p(x \mid y)$ is stable but the marginal $p(y)$ changes across domains. Label shift arises broadly, from clinical deployment across hospital sites to scientific instrument migration and manufacturing process changes, wherever the frequency of outcomes differs between training and test conditions.

Classical label-shift adaptation is usually framed as an estimation problem. In identifiable settings, most notably classification with a well-conditioned source predictor, one can use sufficient *unlabeled* target inputs to estimate the shift ratio $w(y) = p_t(y)/p_s(y)$ without target labels and then retilt the source predictive (Lipton et al., 2018; Alexandari et al., 2020; Garg et al., 2020). This is the right tool when unlabeled target data are available in sufficient quantity and the estimation problem is well conditioned.

The setting considered in this paper is different. In many deployments, *labeled* target outcomes arrive sequentially and may be scarce, unlabeled target inputs alone may be insufficient to estimate the shift reliably, and practitioners already have an externally specified candidate correction, such as a regulatory specification, a bridging experiment, or domain knowledge about the direction and approximate degree of the shift. The question then changes from *"what is the shift?"* to *"is the proposed correction supported by the incoming labeled outcomes?"* This is a confirmation problem, not a shift-estimation problem, and it calls for a different tool.

This paper provides that tool. Given a source predictive $p_s(y \mid x, \mathcal{D}_{\mathrm{tr}})$ and a candidate weight $w(y)$ derived from domain knowledge, we form the tilted predictive $\widetilde{p}_w(y \mid x, \mathcal{D}_{\mathrm{tr}}) \propto w(y)\, p_s(y \mid x, \mathcal{D}_{\mathrm{tr}})$ and accumulate the per-observation likelihood ratio between $\widetilde{p}_w$ and p_s. We show this ratio is a conditional e-value under the predictive null, and its running product is a nonnegative martingale, also known as an e-process (Vovk & Wang, 2021; Shafer, 2021; Ramdas et al., 2023). Ville's inequality then gives an anytime-valid confirmation rule in which Type I error is controlled at every sample size simultaneously, with no correction for optional stopping or repeated looks at the data.

One point deserves emphasis. Rejection at time τ^* means the cumulative negative log-predictive density (NLPD) gap has exceeded $\log(1/\alpha)$, which by Ville's inequality has probability at most α under H_0^{pred}. This confirms the posited correction is consistent with the incoming data, but does not quantify the degree of shift. For quantitative shift estimation, unlabeled-target methods such as BBSE (Lipton et al., 2018) remain the appropriate tool.

This paper contributes (i) the NLPD-gap identity connecting Kelly-optimal betting (Kelly, 1956) to model evalua-

[1]CROID Research and aSSIST University, Seoul, Korea. Correspondence to: Seungjin Choi <seungjin.choi.mlg@gmail.com>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea.
Copyright 2026 by the author(s).

tion; (ii) an anytime-valid confirmation test with no pre-specified sample size; (iii) closed-form GP e-values under Gaussian label-shift tilts; (iv) a growth-rate characterization giving asymptotic power one under correct specification; and (v) empirical demonstrations of Type I control, miscalibration sensitivity, small-batch advantage over noisy label-based re-estimation, and robustness to moderate shift misspecification.

The theory of e-values and e-processes is developed in (Vovk & Wang, 2021; Shafer, 2021; Ramdas et al., 2023). Growth-rate-optimal (GRO) e-values are characterized by Grünwald et al. (2024). Our construction deliberately uses a fixed domain-knowledge-based weight rather than a GRO weight, which would require specifying the alternative more precisely. The same predictive-level tilting construction was used in our companion work on conformal Bayes under label shift (Choi, 2026), where the tilted predictive determines the nonconformity score geometry and is paired with an importance-weighted conformal quantile for target-domain coverage.

In this paper, the same tilted/source predictive pair serves a different purpose. The normalized ratio itself becomes a sequential e-value, yielding anytime-valid confirmation of a fixed posited shift rather than conformal prediction sets. Our test differs from change-point detectors (Vovk et al., 2005; Xu & Xie, 2023) and weighted conformal prediction (Tibshirani et al., 2019; Podkopaev & Ramdas, 2021), which test an unknown alternative or provide coverage guarantees rather than anytime-valid confirmation of a fixed posited shift.

2. Background and Notation

2.1. E-Values, E-Processes, and Ville's Inequality

Let $\mathcal{F}_t = \sigma(\mathcal{D}_{\mathrm{tr}}, X_1, Y_1, \ldots, X_t, Y_t)$ denote the natural filtration (all information up to time t). A nonnegative random variable e is an *e-value* for H_0 if $\mathbb{E}_{H_0}[e] \le 1$. E-values compose sequentially. If $\mathbb{E}_{H_0}[e_i \mid \mathcal{F}_{i-1}] \le 1$ for each i, then $M_t = \prod_{i=1}^{t} e_i$ is a nonnegative supermartingale (an *e-process*) under H_0 (Ramdas et al., 2023); when equality holds it is a martingale. Ville's inequality gives

$$\mathbb{P}_{H_0}\left(\sup_{t\ge 0} M_t > \tfrac{1}{\alpha}\right) \le \alpha,$$

so the stopping rule $\tau^* = \inf\{t : M_t > 1/\alpha\}$ controls Type I error at level α at every sample size simultaneously. This is the *anytime-valid* guarantee. The simplest e-value for testing null $Y_i \mid X_i \sim p_0(\cdot \mid x_i)$ against alternative $Y_i \mid X_i \sim p_1(\cdot \mid x_i)$ is the likelihood ratio $e_i = p_1(Y_i \mid X_i)/p_0(Y_i \mid X_i)$, since $\mathbb{E}_{p_0}[e_i \mid X_i, \mathcal{F}_{i-1}] = 1$.

2.2. Predictive Null, Tilted Predictive, and NLPD

We test the *predictive null*

$$H_0^{\mathrm{pred}} : \; Y_i \mid X_i, \mathcal{F}_{i-1} \sim p_s(\cdot \mid X_i, \mathcal{D}_{\mathrm{tr}}), \tag{1}$$

where $p_s(y \mid x, \mathcal{D}_{\mathrm{tr}})$ is the source posterior predictive, trained on $\mathcal{D}_{\mathrm{tr}}$ and held fixed. This differs from the data-level no-shift null $H_0^{\mathrm{data}} : q_t = q_s$; the two coincide only when p_s is well-calibrated for q_s.

Under label shift $q_s(x \mid y) = q_t(x \mid y)$, so the density ratio $q_t/q_s = q_t(y)/q_s(y) =: w_{\mathrm{true}}(y)$ depends only on y. Given a weight $w(y)$ from domain knowledge, the *tilted predictive* is given by

$$\widetilde{p}_w(y \mid x, \mathcal{D}_{\mathrm{tr}}) := \frac{w(y)\, p_s(y \mid x, \mathcal{D}_{\mathrm{tr}})}{Z_w(x)}, \tag{2}$$

where $Z_w(x) = \int w(y)\, p_s(y \mid x, \mathcal{D}_{\mathrm{tr}})\, dy$. The per-observation e-value is then

$$e_i = \widetilde{p}_w(y_i \mid x_i, \mathcal{D}_{\mathrm{tr}})/p_s(y_i \mid x_i, \mathcal{D}_{\mathrm{tr}}) = w(y_i)/Z_w(x_i).$$

The *negative log-predictive density* (NLPD) of predictive p at (x_i, y_i) is $\mathrm{NLPD}_i(p) := -\log p(y_i \mid x_i, \mathcal{D}_{\mathrm{tr}})$. With $\mathrm{NLPD}_i^s := \mathrm{NLPD}_i(p_s)$ and $\mathrm{NLPD}_i^w := \mathrm{NLPD}_i(\widetilde{p}_w)$ (superscript w for posited weight, not time), $\log e_i = \mathrm{NLPD}_i^s - \mathrm{NLPD}_i^w$, so the log-wealth is the cumulative NLPD gap

$$\log M_t = \sum_{i=1}^{t} (\mathrm{NLPD}_i^s - \mathrm{NLPD}_i^w). \tag{3}$$

3. From Predictive Tilting to Anytime-Valid E-Processes

3.1. E-Value and E-Process Identification

Throughout this subsection the candidate weight w is fixed before the stream begins, $Z_w(x) = \int w(y) p_s(y \mid x, \mathcal{D}_{\mathrm{tr}}) dy$ is finite and positive, and

$$\widetilde{p}_w(y \mid x, \mathcal{D}_{\mathrm{tr}}) = \frac{w(y) p_s(y \mid x, \mathcal{D}_{\mathrm{tr}})}{Z_w(x)}.$$

The inputs X_i may be stochastic, deterministic, or adaptively selected; what matters for validity is the conditional distribution of Y_i under the predictive null.

Lemma 1 (Tilted predictive as likelihood ratio)**.** *For every x with $0 < Z_w(x) < \infty$,*

$$\frac{\widetilde{p}_w(y \mid x, \mathcal{D}_{\mathrm{tr}})}{p_s(y \mid x, \mathcal{D}_{\mathrm{tr}})} = \frac{w(y)}{Z_w(x)}, \qquad \mathbb{E}_{Y \sim p_s(\cdot \mid x, \mathcal{D}_{\mathrm{tr}})}\left[\frac{w(Y)}{Z_w(x)}\right] = 1.$$

Proof. See section A.1.

Proposition 1 (Per-sample predictive e-value)**.** *Define*

$$e_i := \frac{\widetilde{p}_w(Y_i \mid X_i, \mathcal{D}_{\mathrm{tr}})}{p_s(Y_i \mid X_i, \mathcal{D}_{\mathrm{tr}})} = \frac{w(Y_i)}{Z_w(X_i)}.$$

Under H_0^{pred} in (1),

$$\mathbb{E}[e_i \mid \mathcal{F}_{i-1}, X_i] = 1.$$

Hence e_i is an e-value conditionally on the realized input X_i.

Proof. See section A.2.

Proposition 2 (Prequential predictive e-process)**.** *Assume H_0^{pred} and define $M_0 = 1$,*

$$M_t := \prod_{i=1}^{t} e_i.$$

Then $(M_t)_{t\geq 0}$ is a nonnegative martingale with respect to $(\mathcal{F}_t)_{t\geq 0}$. In particular, it is an e-process under the predictive null.

Proof. See section A.3.

By (3), $\log M_t = \sum_{i=1}^{t}(\mathrm{NLPD}_i^s - \mathrm{NLPD}_i^w)$ for every realized stream. Crossing $\log(1/\alpha)$ is therefore exactly the event that the tilted predictive has accumulated more than $\log(1/\alpha)$ nats of log-score advantage over the source. The null assumption enters only when converting this algebraic identity into the anytime-valid test below.

Proposition 3 (Anytime-valid predictive confirmation test)**.** *Under H_0^{pred}, for any $\alpha \in (0,1)$ the stopping rule*

$$\tau^* := \inf\{t \geq 1 : M_t > 1/\alpha\}$$

satisfies

$$\mathbb{P}_{H_0^{\mathrm{pred}}}(\tau^* < \infty) \leq \alpha.$$

Proof. See section A.4.

Remark 1 (What rejection means)**.** Rejecting at time τ^* means the tilted predictive $\widetilde{p}_w$ has outpredicted the source predictive p_s by a cumulative log-score margin exceeding $\log(1/\alpha)$ nats, an event that occurs with probability at most α under H_0^{pred}. This is what is directly confirmed.

When the source predictive p_s is well-calibrated, rejection further implies that the data are inconsistent with no shift and consistent with the posited tilt $w(y)$. It does *not* imply that $w(y)$ is the uniquely correct weight, since a different tilt in the same direction might also have been confirmed. Rejection should therefore be read as *the posited shift correction is supported by the data*, not as *the posited correction is exactly right*.

This distinction matters practically because if the posited shift is in the right direction but the wrong magnitude, the test may still reject, and the practitioner should not interpret rejection as a precise quantification of the shift. For that, estimation methods (Lipton et al., 2018; Alexandari et al., 2020) are the right tool. The confirmation test answers the binary question "Is the correction plausible?" rather than a quantitative one.

Proposition 4 (Growth rate and asymptotic power)**.** *Condition on the fixed training data $\mathcal{D}_{\mathrm{tr}}$. Suppose $(X_i, Y_i) \overset{\mathrm{iid}}{\sim} \nu(dx)\,\widetilde{p}_w(dy \mid x, \mathcal{D}_{\mathrm{tr}})$, i.e. inputs are drawn i.i.d. from a marginal ν and, given $X_i = x$, outcomes follow the tilted predictive. If*

$$\gamma := \mathbb{E}_{X\sim\nu}\big[D_{\mathrm{KL}}\big(\widetilde{p}_w(\cdot \mid X, \mathcal{D}_{\mathrm{tr}}) \,\|\, p_s(\cdot \mid X, \mathcal{D}_{\mathrm{tr}})\big)\big] > 0$$

and $\mathbb{E}|\log e_i| < \infty$, then

$$\frac{1}{t}\log M_t \to \gamma \quad \textit{a.s.}, \qquad \mathbb{P}_{\widetilde{p}_w}(\tau^* < \infty) = 1.$$

Proof. See section A.5.

Proposition 4 has two complementary consequences. First, the growth rate γ equals the average KL divergence between the tilted and source predictives. Since $D_{\mathrm{KL}}(\widetilde{p}_w(\cdot \mid X, \mathcal{D}_{\mathrm{tr}}) \| p_s(\cdot \mid X, \mathcal{D}_{\mathrm{tr}})) \geq 0$, the growth rate is positive whenever $\widetilde{p}_w(\cdot \mid X, \mathcal{D}_{\mathrm{tr}}) \neq p_s(\cdot \mid X, \mathcal{D}_{\mathrm{tr}})$ on a set of positive ν-measure. In that case $\mathbb{P}_{\widetilde{p}_w}(\tau^* < \infty) = 1$, so the test eventually rejects H_0^{pred} with probability one. Second, γ quantifies the speed. A larger shift (higher KL) gives faster rejection, with expected stopping time approximately $\log(1/\alpha)/\gamma$. This tells the practitioner how many target observations to expect before confirmation, purely from the posited shift and the source predictive, with no target data required. Together, Proposition 3 (the test is safe) and Proposition 4 (the test is powerful, and quantifiably so under correct specification) give a model-based frequentist characterization of the confirmation procedure.

Remark 2 (What happens under misspecification?)**.** If the true conditional distribution is $q_t(\cdot \mid x)$ rather than $\widetilde{p}_w$ or p_s, the expected log-growth is the cross-entropy gap

$$\mathbb{E}_{q_t}\left[\log \widetilde{p}_w(Y \mid X, \mathcal{D}_{\mathrm{tr}}) - \log p_s(Y \mid X, \mathcal{D}_{\mathrm{tr}})\right].$$

This quantity is positive exactly when the tilted predictive improves the expected log score relative to the source predictive. However, exact Type I control under the data-level no-shift null $q_t = q_s$ is guaranteed only if $q_s(Y \mid X) = p_s(Y \mid X, \mathcal{D}_{\mathrm{tr}})$, or more generally if $\mathbb{E}_{q_s}[e_i \mid X_i, \mathcal{F}_{i-1}] \leq 1$. Thus the present test is safe relative to a fixed predictive null; robustness to arbitrary predictive misspecification would require an additional calibration or conformalization layer.

3.2. Closed-Form E-Values for GP Sources

Assumption 1 (GP source and Gaussian posited tilt)**.** $p_s(y \mid x, \mathcal{D}_{\mathrm{tr}}) = \mathcal{N}(y; \mu_s(x), \sigma_s^2(x))$. *The posited source*

and target label marginals used to form the weight are $p_s^{\mathrm{lab}}(y) = \mathcal{N}(y; m_s, v_s^2)$ *and* $p_t^{\mathrm{lab}}(y) = \mathcal{N}(y; m_t, v_t^2)$*, with* $v_s, v_t > 0$*. For every tested input* x*,*

$$\Lambda(x) := \frac{1}{\sigma_s^2(x)} + \frac{1}{v_t^2} - \frac{1}{v_s^2} > 0.$$

Remark 3 (Positive-precision condition in practice)**.** The necessary and sufficient condition for the Gaussian normalization below is $\Lambda(x) > 0$. The simpler condition $\sigma_s^2(x) < v_t^2$ is neither necessary nor sufficient in general. If $v_t^2 \le v_s^2$, then $\Lambda(x) > 0$ automatically for every $\sigma_s^2(x) > 0$. If $v_t^2 > v_s^2$, then the condition becomes $\sigma_s^2(x) < v_s^2 v_t^2/(v_t^2 - v_s^2)$. When $\Lambda(x) \le 0$, the untruncated Gaussian-ratio weight does not define a finite normalizer at that input; the closed form is not merely unavailable. One must either restrict testing to inputs satisfying $\Lambda(x) > 0$, modify or truncate the posited weight so that $Z_w(x) < \infty$, or use a different predictive/weight pair for which the normalizer is finite. Numerical quadrature is appropriate only after such finiteness has been guaranteed. The e-process property is preserved exactly whenever the same finite $Z_w(x)$ is used in both $\widetilde{p}_w$ and e_i.

Under assumption 1, the posited weight $w(y) = (v_s/v_t)\exp\big((y - m_s)^2/(2v_s^2) - (y - m_t)^2/(2v_t^2)\big)$ is exponential-quadratic.

Proposition 5 (Closed-form normalization)**.** *Under assumption 1, with* $\Lambda(x) := 1/\sigma_s^2(x) + 1/v_t^2 - 1/v_s^2$ *and* $\mu^*(x) := \Lambda(x)^{-1}(\mu_s(x)/\sigma_s^2(x) + m_t/v_t^2 - m_s/v_s^2)$*,*

$$Z_w(x) = \eta(x)\exp\Big(\tfrac{(\mu^*(x))^2\Lambda(x)}{2} - \tfrac{\mu_s^2(x)}{2\sigma_s^2(x)} - \tfrac{m_t^2}{2v_t^2} + \tfrac{m_s^2}{2v_s^2}\Big),$$

where $\eta(x) = \frac{v_s}{v_t\,\sigma_s(x)\sqrt{\Lambda(x)}}$*.*

Proof. See section A.6.

Proposition 6 (Closed-form per-sample log e-value)**.** *Under assumption 1, the tilted predictive is* $\widetilde{p}_w(\cdot \mid x, \mathcal{D}_{\mathrm{tr}}) = \mathcal{N}(\mu^*(x), \Lambda(x)^{-1})$*, and*

$$\log e_i = \frac{(y_i - \mu_s(x_i))^2}{2\sigma_s^2(x_i)} - \frac{\Lambda(x_i)}{2}(y_i - \mu^*(x_i))^2 + \frac{1}{2}\log\big(\Lambda(x_i)\,\sigma_s^2(x_i)\big),$$

computable from GP posterior outputs $(\mu_s(x_i), \sigma_s^2(x_i))$ *and the shift parameters* (m_s, v_s^2, m_t, v_t^2) *alone.*

Proof. See section A.7.

Remark 4 ($p(x)$ cancellation)**.** Under label shift, $w(y)$ does not depend on x, so $p_s(x)$ cancels from all expressions. proposition 6 inherits this because $\log e_i$ depends on x_i only through GP posterior outputs $\mu_s(x_i)$ and $\sigma_s^2(x_i)$, requiring no knowledge of the feature distribution.

3.3. Application to FDR Control and Robustness Sets

When K streams are tested in parallel, applying the test K times at level α inflates the expected number of false confirmations to $K\alpha$. Bonferroni correction is too conservative (it targets family-wise error). Instead, the e-BH procedure of Wang & Ramdas (2022) controls the *false discovery rate* (FDR), the expected fraction of false confirmations among all confirmed streams, at level α. For each compound k, compute the batch e-value

$$e^{(k)} := M_{n_k}^{(k)} = \exp\Big(\sum_{i=1}^{n_k}\big(\mathrm{NLPD}_i^{s,(k)} - \mathrm{NLPD}_i^{w,(k)}\big)\Big),$$

which is valid under $H_0^{\mathrm{pred},(k)}$ by proposition 2. Ordering $e^{(1)} \ge \cdots \ge e^{(K)}$ and setting $k^* = \max\{k : e^{(k)} \ge K/(\alpha k)\}$, rejecting the top-$k^*$ controls $\mathrm{FDR} \le \alpha$ under independence across streams. The threshold $K/(\alpha k)$ relaxes progressively with rank, giving e-BH more power than Bonferroni. We recommend the batch e-value as the e-BH input since it is order-free; using a sequentially stopped value $M_{\tau^*}^{(k)}$ is valid (Ramdas et al., 2023) but requires a careful per-compound protocol.

If the practitioner is uncertain about the exact weight, any fixed $w \in \mathcal{W}_\beta$ from a robustness set (e.g. a χ^2 ball) gives a valid e-process by proposition 2. A mixture e-process $M_t^{\mathrm{mix}} = \int M_t(w)\,d\Pi(w)$ over a prior Π chosen before testing is also valid; designing growth-rate-optimal safe tests for the composite alternative connects to GRO e-values (Grünwald et al., 2024) and is left to future work.

4. Confirmation Test in Practice

The method requires a strict two-stage separation. In the first stage, the practitioner fixes the weight $w(y)$ using any available information *before* the testing stream begins. This could be an off-the-shelf label-shift estimator such as BBSE (Lipton et al., 2018) or MLLS (Alexandari et al., 2020) applied to a *separate* pilot dataset, domain expert knowledge, a regulatory specification, or a bridging experiment. In the second stage, the fixed w is handed to Algorithm 1, which accumulates sequential evidence for or against it on the incoming target stream. The validity guarantee (proposition 3) depends critically on this separation because w must not depend on the observations being tested, since the e-value $e_i = w(y_i)/Z_w(x_i)$ is a valid e-value only when w is committed to before y_i is observed. If w was estimated on an independent pilot batch, that batch must be disjoint from the testing stream.

Algorithm 1 summarizes the complete procedure. The offline phase fixes the source predictive and the posited shift parameters; it requires no additional estimation from the target testing stream. The online phase adds only constant-time arithmetic after the GP predictive $(\mu_s(x_i), \sigma_s^2(x_i))$ has

Algorithm 1: Anytime-Valid Predictive Confirmation Test for a Posited Label-Shift Tilt

Inputs : Source GP outputs $\mu_s(\cdot), \sigma_s^2(\cdot)$ from $p_s(y \mid x, \mathcal{D}_{\mathrm{tr}})$
Posited shift parameters (m_s, v_s^2, m_t, v_t^2)
Significance level $\alpha \in (0,1)$; optional monitoring horizon T
Output : Rejection time τ^*, or "not rejected" after T observations

```
// Offline stage. Commit to the source predictive and the posited tilt
```
Fix (m_s, v_s^2, m_t, v_t^2) and set log-wealth $S_0 \leftarrow 0$
```
// Online stage. Process target observations sequentially
```
for $i = 1, 2, \ldots, T$ **do**
  Receive (x_i, y_i)
  Evaluate $\Lambda_i \leftarrow 1/\sigma_s^2(x_i) + 1/v_t^2 - 1/v_s^2$ and $\mu_i^* \leftarrow \Lambda_i^{-1}(\mu_s(x_i)/\sigma_s^2(x_i) + m_t/v_t^2 - m_s/v_s^2)$
  if $\Lambda_i \leq 0$ **then**
    return "invalid Gaussian tilt at x_i; modify or truncate w so that $Z_w(x_i) < \infty$"
  end
  Compute log e-value using proposition 6 $\log e_i \leftarrow \dfrac{(y_i - \mu_s(x_i))^2}{2\sigma_s^2(x_i)} - \dfrac{\Lambda_i}{2}(y_i - \mu_i^*)^2 + \dfrac{1}{2}\log\big(\Lambda_i\, \sigma_s^2(x_i)\big)$
  Update log-wealth, $S_i \leftarrow S_{i-1} + \log e_i$ // $S_i = \log M_i$ by (3)
  if $S_i > \log(1/\alpha)$ **then return** $\tau^* \leftarrow i$ // tilted predictive confirmed (proposition 3)
end
return "not rejected after T observations"

been evaluated. The method computes $\log e_i$ via proposition 6 and increments the running log-wealth S_t. By (3), $S_t = \log M_t$ is exactly the cumulative NLPD gap, so practitioners already tracking that gap are implicitly running the test.

5. Experiments

We illustrate the method through end-to-end GP regression simulations that follow the intended two-stage workflow. We train a source GP on source data, specify a label-shift correction, and accumulate closed-form e-values on target observations. This design tests the full pipeline, including GP posterior uncertainty propagation into the e-values.

In each trial we generate a synthetic source training set

$$y_j^{\mathrm{tr}} = f(x_j^{\mathrm{tr}}) + \epsilon_j,$$

where $x_j^{\mathrm{tr}} \sim \mathrm{Unif}[-3, 3]$ and $\epsilon_j \sim \mathcal{N}(0, \sigma_\epsilon^2)$, with $f(x) = \sin(2x) + 0.5\cos(4x)$. We fit a GP with an RBF kernel to $\mathcal{D}_{\mathrm{tr}} = \{(x_j^{\mathrm{tr}}, y_j^{\mathrm{tr}})\}_{j=1}^{n_{\mathrm{tr}}}$. The GP posterior predictive

$$p_s(y \mid x, \mathcal{D}_{\mathrm{tr}}) = \mathcal{N}(y; \mu_s(x), \sigma_s^2(x))$$

serves as the source predictive. Unless stated otherwise, $n_{\mathrm{tr}} = 50$, $\sigma_\epsilon^2 = 1$, RBF amplitude 1 and lengthscale 0.8, $\alpha = 0.05$. The posited label marginals are $p_s^{\mathrm{lab}}(y) = \mathcal{N}(y; 0, 2)$ and $p_t^{\mathrm{lab}}(y) = \mathcal{N}(y; m_t, 3)$, and shift magnitude is $\Delta = (m_t - m_s)/\sqrt{v_s^2}$. Under H_0^{pred}, target labels are drawn from $p_s(\cdot \mid x, \mathcal{D}_{\mathrm{tr}})$; under the alternative, from $\widetilde{p}_w(\cdot \mid x, \mathcal{D}_{\mathrm{tr}})$.

Experiment 1. E-process trajectories. Fig. 1 shows median $\log M_t$ with IQR bands across 500 trials for one null stream and for alternatives with $\Delta \in \{0.5, 1.0, 1.5\}$. Under H_0^{pred}, log-wealth fluctuates around zero and stays well below the rejection boundary $\log(1/\alpha) \approx 3$. Under the alternative, log-wealth grows approximately linearly and the empirical slope closely tracks the conditional KL rate γ predicted by proposition 4; the slope increases with Δ. This confirms that the growth-rate characterization holds in the end-to-end GP setting, where $\mu_s(x)$ and $\sigma_s^2(x)$ come from a fitted model rather than from a theoretical oracle.

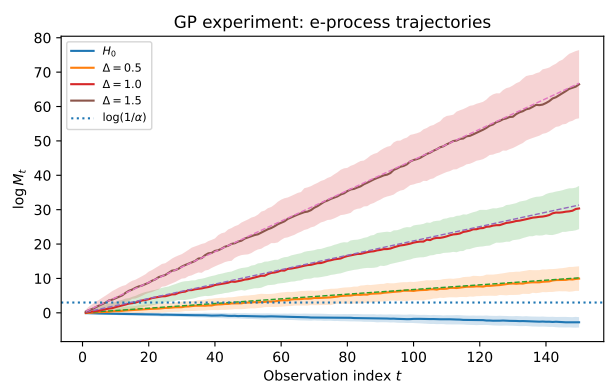

Figure 1. E-process trajectories under end-to-end GP regression ($n_{\mathrm{tr}} = 50$, 500 trials). Each trial trains a GP on source data and computes closed-form e-values from the GP posterior. Dashed lines show the KL growth rate predicted by proposition 4; the dotted line is $\log(1/\alpha)$.

Experiment 2. Type I error under the predictive null. Fig. 2 plots the empirical false-alarm probability $\widehat{\alpha}(T) = \mathbb{P}_{H_0^{\mathrm{pred}}}(\tau^* \leq T)$ over 2000 trials when target labels are drawn from the fitted source GP predictive. The curve stays strictly below $\alpha = 0.05$ at every displayed horizon, confirming the anytime-valid guarantee of proposition 3 in the end-to-end GP setting. The overall false-alarm rate at $T = 200$ is 3.25%, demonstrating that the closed-form e-values are conservative in this GP regression setup.

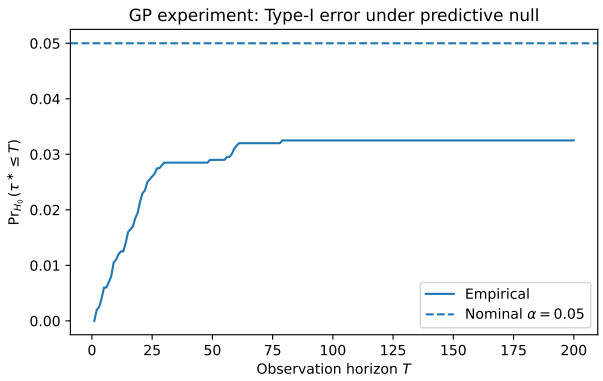

Figure 2. Type I error under H_0^{pred} ($n_{\text{tr}} = 50$, 2000 trials). The empirical false-alarm probability $\widehat{\alpha}(T)$ stays below $\alpha = 0.05$ (dashed) at every horizon, confirming the anytime-valid guarantee with a fitted GP source predictive.

Experiment 3. Power vs fixed-sample LRT. Fig. 3 compares the anytime-valid e-process with a fixed-sample likelihood-ratio test (LRT) calibrated separately at each horizon under H_0^{pred}, at $\Delta = 1.0$ (1500 trials). The fixed-sample LRT knows the sample size in advance and is optimally calibrated at each horizon; the e-process does not. Accordingly, the LRT has a power advantage at large horizons. At $n = 50$, power is 0.993 for the LRT and 0.963 for the e-process. However, the e-process provides validity under *arbitrary* stopping. Its α guarantee holds regardless of when the practitioner stops, with no recalibration. The LRT, by contrast, is invalid if the practitioner stops early or peeks at interim results. This is the fundamental trade-off. The e-process sacrifices a small amount of power at any fixed horizon in exchange for the freedom to stop at any time.

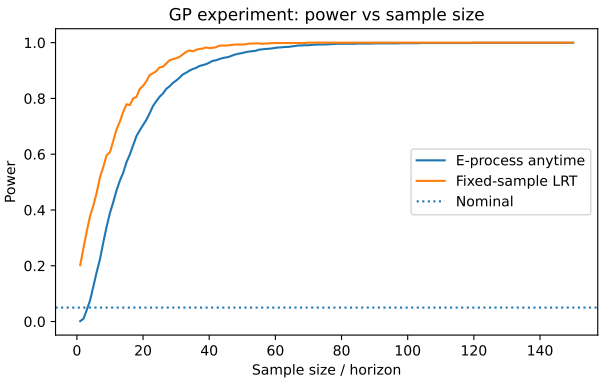

Figure 3. Power vs sample size at $\Delta = 1.0$ ($n_{\text{tr}} = 50$, 1500 trials). The fixed-sample LRT is calibrated separately at each horizon and is valid only at the pre-specified n; the e-process is valid under any stopping time. The power gap reflects the cost of anytime-validity.

Experiment 4. Reliable prior shift vs adaptive re-estimation. For each target batch size n_t, we compare three predictives on a held-out target set. The three predictives are (i) the uncorrected source GP, (ii) the oracle corrected predictive using the true shift weight, and (iii) an adaptive correction that estimates m_t from the n_t target observations and applies the plug-in tilt. Fig. 4 shows that the oracle correction gives substantially lower held-out NLPD than the source predictive at all n_t, while re-estimation is competitive only for larger batches. At $n_t = 10$, mean NLPD is 1.794 (source), 1.586 (oracle correction), and 1.652 (re-estimated), confirming the small-batch advantage of a reliable prior specification. This motivates the confirmation test. When a trustworthy prior tilt is available, using it beats noisy re-estimation, and the confirmation test provides formal evidence that the tilt is consistent with the data.

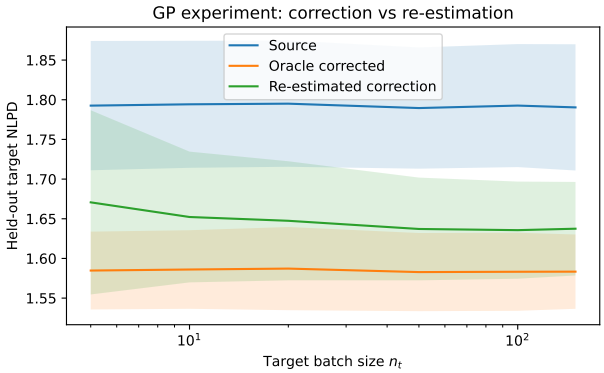

Figure 4. Oracle correction vs adaptive re-estimation under end-to-end GP regression. Held-out target NLPD vs target batch size n_t. A reliable prior tilt dominates at small n_t; re-estimation improves as n_t grows but remains noisier in the small-batch regime.

Experiment 5. Source miscalibration stress test. The predictive-null guarantee requires observations to follow the source predictive. To stress-test this, we generate null labels from $\mathcal{N}(\mu_s(x), c\,\sigma_s^2(x))$ but compute e-values using the original GP predictive $\mathcal{N}(\mu_s(x), \sigma_s^2(x))$, simulating a GP whose posterior variance is off by a factor c. Fig. 5 shows that at $c = 1$ (correct calibration) Type I error is 3.2%, well below $\alpha = 0.05$. Variance underestimation ($c > 1$) is severely anti-conservative. Type I error rises to 36.45% at $c = 2$ and 90.65% at $c = 4$. Variance overestimation ($c < 1$) is conservative. This confirms that the predictive-null caveat is operationally significant because a poorly calibrated GP posterior can substantially inflate false-alarm rates, and practitioners should verify GP calibration before deployment.

Order-sensitivity and batch-e-value comparison experiments are omitted for space; results are consistent with the theoretical analysis in remark 2.

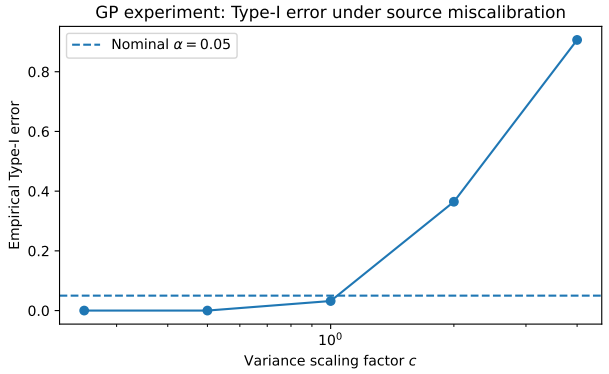

Figure 5. Type I error under GP posterior miscalibration. Null labels are drawn from $\mathcal{N}(\mu_s(x), c\sigma_s^2(x))$ but e-values use the original GP predictive ($c = 1$). Variance underestimation ($c > 1$) inflates Type I error sharply; $c = 1$ is calibrated, while $c < 1$ is conservative.

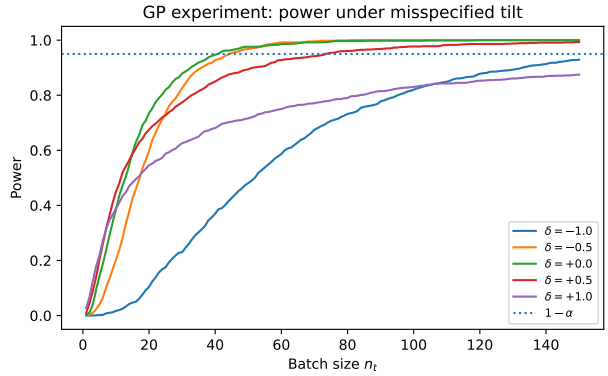

Figure 6. Power under misspecified tilt ($n_{\text{tr}} = 50$, 1000 trials, $\Delta = 1.0$). The true target is generated from the correctly tilted predictive; the test uses a posited mean offset by δ. Moderate misspecification reduces power but does not break the test.

Experiment 6. Misspecified tilt sensitivity. The previous experiments assume the posited shift weight is correctly specified. Here we ask what happens when the practitioner's prior is wrong. We generate target labels from the correctly tilted predictive at $\Delta = 1.0$, but compute e-values using a posited target mean offset by $\delta \in \{-1.0, -0.5, 0, +0.5, +1.0\}$ (in units of $\sqrt{v_s^2}$, 1000 trials). Fig. 6 shows power curves and Fig. 7 shows mean $\log M_{50}$.

Two findings confirm the theoretical asymmetry. First, moderate misspecification reduces power, especially when the posited shift undershoots the true shift, but the test remains informative and power improves with batch size. Second, and more importantly, tilt misspecification cannot inflate Type I error under the predictive null as long as the weight is fixed before testing and $Z_w(x_i)$ is finite. propositions 1 and 2 still give a valid e-process. This asymmetry is the key practical reassurance for practitioners who must specify $w(y)$ from imperfect prior knowledge. Validity is robust to the posited tilt, while power degrades with the magnitude and direction of the error.

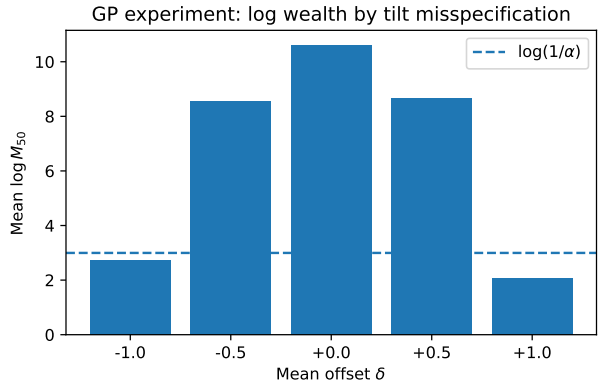

Figure 7. Mean $\log M_{50}$ under misspecified tilt. The dashed line is the rejection threshold $\log(1/\alpha)$. All offsets δ yield positive mean log-wealth, although severe overstatement of the shift may fall below the rejection threshold at $n_t = 50$.

6. Conclusion

When labeled target outcomes are scarce and unlabeled target data alone are insufficient for reliable shift estimation, re-estimating a label shift from scratch is unreliable. This paper provides a formal tool for the complementary setting in which the practitioner has a domain-knowledge-based prior on the shift direction and needs statistical evidence that it is consistent with the incoming data. We showed that the cumulative NLPD gap between a fixed source predictive and a pre-specified tilted predictive is the logarithm of a valid e-process under the source predictive null. The normalized tilted/source likelihood ratios are conditional e-values, their product is a nonnegative martingale, and Ville's inequality yields an anytime-valid confirmation rule with no pre-specified sample size. For GP sources with Gaussian label-shift tilts, e-values are available in closed form. Rejection confirms that the posited shift is directionally supported by the data, not that it is exactly correct; for quantitative shift estimation from target covariates, unlabeled-target adaptation methods such as BBSE remain the appropriate tool.

The main limitation is predictive-null validity. Type I control holds when the source predictive is correctly calibrated, not under arbitrary model misspecification. Experiment 5 shows this is operationally significant. Variance underestimation by a factor of 2 inflates the false-alarm rate to 36.45%. Practitioners should verify GP calibration before deployment. The candidate weight must also be fixed before the testing stream begins. Extensions to distribution-free validity via conformalization, and to composite or misspecified-shift

families via mixture e-processes (Grünwald et al., 2024), are natural future directions.

References

Alexandari, A. M., Kundaje, A., and Shrikumar, A. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Choi, S. Conformal Bayes under label shift: Post-hoc calibration vs. in-training adaptation. In *The 2nd Workshop on Epistemic Intelligence in Machine Learning*, 2026.

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. C. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Grünwald, P., de Heide, R., and Koolen, W. Safe testing. *Journal of the Royal Statistical Society Series B*, 86(5): 1091–1128, 2024.

Kelly, J. L. A new interpretation of information rate. *The Bell System Technical Journal*, 35(4):917–926, 1956.

Lipton, Z. C., Wang, Y.-X., and Smola, A. J. Detecting and correcting for label shift with black box predictors. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Podkopaev, A. and Ramdas, A. Distribution-free uncertainty quantification for classification under label shift. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Shafer, G. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A*, 184(2):407–431, 2021.

Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Vovk, V. and Wang, R. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer, 2005.

Wang, R. and Ramdas, A. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B*, 84(3):822–852, 2022.

Xu, C. and Xie, Y. Sequential predictive conformal inference for time series. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

A. Proofs

A.1. Proof of lemma 1

Proof. For fixed x with $0 < Z_w(x) < \infty$, the tilted predictive satisfies

$$\mathbb{P}_w(dy \mid x, \mathcal{D}_{\mathrm{tr}}) = \frac{w(y)}{Z_w(x)} \mathbb{P}_s(dy \mid x, \mathcal{D}_{\mathrm{tr}}).$$

Thus $\mathbb{P}_w(\cdot \mid x, \mathcal{D}_{\mathrm{tr}})$ is absolutely continuous with respect to $\mathbb{P}_s(\cdot \mid x, \mathcal{D}_{\mathrm{tr}})$ and its Radon–Nikodym derivative is

$$\frac{d\mathbb{P}_w(\cdot \mid x, \mathcal{D}_{\mathrm{tr}})}{d\mathbb{P}_s(\cdot \mid x, \mathcal{D}_{\mathrm{tr}})}(y) = \frac{w(y)}{Z_w(x)} \qquad \mathbb{P}_s(\cdot \mid x, \mathcal{D}_{\mathrm{tr}})\text{-a.s.}$$

In density notation this is

$$\frac{\widetilde{p}_w(y \mid x, \mathcal{D}_{\mathrm{tr}})}{p_s(y \mid x, \mathcal{D}_{\mathrm{tr}})} = \frac{w(y)}{Z_w(x)}$$

wherever the density ratio is defined. Finally,

$$\mathbb{E}_{Y \sim p_s(\cdot \mid x, \mathcal{D}_{\mathrm{tr}})} \left[\frac{w(Y)}{Z_w(x)} \right] = \frac{1}{Z_w(x)} \int w(y) p_s(y \mid x, \mathcal{D}_{\mathrm{tr}}) dy = 1,$$

by the definition of $Z_w(x)$. □

A.2. Proof of proposition 1

Proof. Condition on $(\mathcal{F}_{i-1}, X_i)$. Under H_0^{pred}, the conditional distribution of Y_i is $p_s(\cdot \mid X_i, \mathcal{D}_{\mathrm{tr}})$. Therefore, by lemma 1,

$$\mathbb{E}[e_i \mid \mathcal{F}_{i-1}, X_i] = \mathbb{E}_{Y \sim p_s(\cdot \mid X_i, \mathcal{D}_{\mathrm{tr}})} \left[\frac{w(Y)}{Z_w(X_i)} \right] = 1.$$

Since $e_i \geq 0$, this proves that e_i is a conditional e-value given the realized input X_i. □

A.3. Proof of proposition 2

Proof. Non-negativity is immediate because each $e_i \geq 0$. Moreover, by proposition 1,

$$\mathbb{E}[e_t \mid \mathcal{F}_{t-1}, X_t] = 1.$$

Thus e_t is conditionally integrable, and by the tower property,

$$\mathbb{E}[e_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[\mathbb{E}[e_t \mid \mathcal{F}_{t-1}, X_t] \mid \mathcal{F}_{t-1}] = \mathbb{E}[1 \mid \mathcal{F}_{t-1}] = 1.$$

Since M_{t-1} is $\mathcal{F}_{t-1}$-measurable,

$$\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[M_{t-1} e_t \mid \mathcal{F}_{t-1}] = M_{t-1} \mathbb{E}[e_t \mid \mathcal{F}_{t-1}] = M_{t-1}.$$

Therefore $(M_t)_{t \geq 0}$ is a nonnegative martingale with $M_0 = 1$, and hence an e-process under H_0^{pred}. □

A.4. Proof of proposition 3

Proof. By proposition 2, $(M_t)_{t \geq 0}$ is a nonnegative martingale under H_0^{pred} with $M_0 = 1$. Ville's inequality gives

$$\mathbb{P}_{H_0^{\mathrm{pred}}} \left(\sup_{t \geq 0} M_t > \frac{1}{\alpha} \right) \leq \alpha.$$

The event $\{\tau^* < \infty\}$ is contained in the event $\{\sup_{t \geq 0} M_t > 1/\alpha\}$, so

$$\mathbb{P}_{H_0^{\mathrm{pred}}}(\tau^* < \infty) \leq \alpha.$$

□

A.5. Proof of proposition 4

Proof. Condition on the fixed training data $\mathcal{D}_{\mathrm{tr}}$. Under the assumption $(X_i, Y_i) \overset{\mathrm{iid}}{\sim} \nu(dx)\widetilde{p}_w(dy \mid x, \mathcal{D}_{\mathrm{tr}})$, the variables $\log e_i$ are i.i.d. and integrable. For any fixed x,

$$\begin{aligned}\mathbb{E}_{\widetilde{p}_w(\cdot \mid x, \mathcal{D}_{\mathrm{tr}})}[\log e_i] &= \int \widetilde{p}_w(y \mid x, \mathcal{D}_{\mathrm{tr}}) \log \frac{\widetilde{p}_w(y \mid x, \mathcal{D}_{\mathrm{tr}})}{p_s(y \mid x, \mathcal{D}_{\mathrm{tr}})} dy \\ &= D_{\mathrm{KL}}\{\widetilde{p}_w(\cdot \mid x, \mathcal{D}_{\mathrm{tr}}) \| p_s(\cdot \mid x, \mathcal{D}_{\mathrm{tr}})\}.\end{aligned}$$

Taking expectation over $X \sim \nu$ gives

$$\mathbb{E}[\log e_i] = \mathbb{E}_{X \sim \nu}\left[D_{\mathrm{KL}}\{\widetilde{p}_w(\cdot \mid X, \mathcal{D}_{\mathrm{tr}}) \| p_s(\cdot \mid X, \mathcal{D}_{\mathrm{tr}})\}\right] = \gamma.$$

The strong law of large numbers yields

$$\frac{1}{t} \log M_t = \frac{1}{t} \sum_{i=1}^{t} \log e_i \to \gamma \qquad \text{a.s.}$$

If $\gamma > 0$, then $\log M_t \to \infty$ almost surely. Hence the fixed threshold $\log(1/\alpha)$ is eventually crossed with probability one, so $\mathbb{P}_{\widetilde{p}_w}(\tau^* < \infty) = 1$. □

A.6. Proof of proposition 5

Proof. We compute $Z_w(x) = \int w(y) p_s(y \mid x, \mathcal{D}_{\mathrm{tr}}) dy$ by substituting assumption 1.

$$Z_w(x) = \int_{-\infty}^{\infty} \frac{v_s}{v_t} \exp\left(\frac{(y-m_s)^2}{2v_s^2} - \frac{(y-m_t)^2}{2v_t^2}\right) \frac{1}{\sqrt{2\pi}\sigma_s(x)} \exp\left(-\frac{(y-\mu_s(x))^2}{2\sigma_s^2(x)}\right) dy.$$

The exponent can be written as

$$-\frac{\Lambda(x)}{2} y^2 + y\left\{\frac{\mu_s(x)}{\sigma_s^2(x)} + \frac{m_t}{v_t^2} - \frac{m_s}{v_s^2}\right\} + C_0,$$

where

$$C_0 = -\frac{\mu_s^2(x)}{2\sigma_s^2(x)} - \frac{m_t^2}{2v_t^2} + \frac{m_s^2}{2v_s^2}.$$

By definition,

$$\mu^*(x) = \Lambda(x)^{-1}\left\{\frac{\mu_s(x)}{\sigma_s^2(x)} + \frac{m_t}{v_t^2} - \frac{m_s}{v_s^2}\right\},$$

so the exponent is

$$-\frac{\Lambda(x)}{2}\{y - \mu^*(x)\}^2 + C_0 + \frac{\Lambda(x)(\mu^*(x))^2}{2}.$$

Since $\Lambda(x) > 0$ by assumption 1,

$$\int_{-\infty}^{\infty} \exp\left[-\frac{\Lambda(x)}{2}\{y - \mu^*(x)\}^2\right] dy = \sqrt{\frac{2\pi}{\Lambda(x)}}.$$

Combining the prefactors gives

$$Z_w(x) = \frac{v_s}{v_t \sigma_s(x)\sqrt{\Lambda(x)}} \exp\left(\frac{\Lambda(x)(\mu^*(x))^2}{2} - \frac{\mu_s^2(x)}{2\sigma_s^2(x)} - \frac{m_t^2}{2v_t^2} + \frac{m_s^2}{2v_s^2}\right),$$

which is the stated expression. □

A.7. Proof of proposition 6

Proof. The completing-the-square calculation in the proof of proposition 5 shows that the unnormalized tilted density is proportional to

$$\exp\left[-\frac{\Lambda(x)}{2}\{y-\mu^*(x)\}^2\right].$$

After normalization, therefore,

$$\widetilde{p}_w(\cdot \mid x, \mathcal{D}_{\mathrm{tr}}) = \mathcal{N}(\mu^*(x), \Lambda(x)^{-1}).$$

At (x_i, y_i),

$$\log \widetilde{p}_w(y_i \mid x_i, \mathcal{D}_{\mathrm{tr}}) = -\frac{\Lambda(x_i)}{2}\{y_i-\mu^*(x_i)\}^2 - \frac{1}{2}\log\{2\pi\Lambda(x_i)^{-1}\},$$
$$\log p_s(y_i \mid x_i, \mathcal{D}_{\mathrm{tr}}) = -\frac{\{y_i-\mu_s(x_i)\}^2}{2\sigma_s^2(x_i)} - \frac{1}{2}\log\{2\pi\sigma_s^2(x_i)\}.$$

Subtracting the second display from the first yields

$$\log e_i = \frac{\{y_i-\mu_s(x_i)\}^2}{2\sigma_s^2(x_i)} - \frac{\Lambda(x_i)}{2}\{y_i-\mu^*(x_i)\}^2 + \frac{1}{2}\log\{\Lambda(x_i)\sigma_s^2(x_i)\},$$

which proves the stated formula. □