DECISION-FOCUSED CALIBRATION WITH ADAPTIVE CONFORMAL PREDICTIVE DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Decision-focused learning can improve downstream decisions, but a predictive distribution trained for a particular task may misrepresent the outcomes it is meant to predict. Conversely, a distribution trained solely for predictive accuracy may miss features that matter most to the decision. We study how to adapt predictive distributions to decisions subject to a conformal calibration constraint. Our key observation is that a nonconformity score separates the remaining modeling problem into two parts: how probability varies across score levels with the context, and which outcomes occur within each level. We propose `AdaCF`, an adaptive conformal flow that learns both parts, optionally incorporates a downstream decision loss, and uses a rank-based correction to ensure finite-sample marginal validity under exchangeability. We decompose predictive error into the errors of the two learned components and bound regret for bounded decision losses in terms of that error. On synthetic problems and electricity price data, `AdaCF` improves predictive fidelity and decision quality over standard conformal flows while preserving validity.

1 INTRODUCTION

Decisions under uncertainty depend on predictions of outcomes that have yet to be observed. An investor allocates a portfolio before returns are realized (Rockafellar & Uryasev, 2000; Butler & Kwon, 2023), a clinician chooses a treatment before observing a patient's response (Qian & Murphy, 2011; Pedone et al., 2024), and an energy operator schedules generation before demand and prices are known (Donti et al., 2017; Ellinas et al., 2024). A common approach is to learn a predictive distribution and optimize a decision against it. Yet a distribution that predicts well on average can be inaccurate where the decision is most sensitive, leading to substantial regret. Decision-focused learning (DFL) addresses this problem by training the predictor using downstream decision loss (Hu et al., 2022; Elmachtoub et al., 2023). At the other extreme, a predictor trained only to obtain a good decision may assign probabilities that poorly reflect the outcomes it purports to predict. Such a model can perform well on the task used for training while offering an unreliable account of uncertainty for other decisions.

These two objectives are often combined through a weighted sum of predictive and decision losses (Wang et al., 2026; Chen et al., 2025). The weight controls a consequential trade-off: how far may the predictive distribution move from the data to improve a particular decision? Its choice is generally task dependent, and a favorable decision loss alone provides no assurance that the resulting probabilities remain meaningful. We instead ask whether a predictive distribution can be adapted to downstream decisions while satisfying an explicit calibration requirement. This leads to a constrained view of decision-focused learning: seek a distribution that supports good decisions and represents the data, subject to a statistical validity condition. Calibration does not by itself identify the true distribution; its role is to impose a verifiable constraint, while predictive fitting determines how probability is allocated within that constraint.

Conformal prediction offers a natural starting point. Under exchangeability, it constructs prediction sets with finite-sample marginal coverage without requiring a correctly specified predictive model (Vovk et al., 2005; Shafer & Vovk, 2008). Sets alone, however, do not assign probabilities to the outcomes they contain, whereas stochastic optimization requires a predictive distribution. Conformal flows turn nested conformal sets into a sampleable distribution whose draws satisfy the corresponding set-coverage requirements (Harris, 2026). This validity condition still leaves considerable freedom.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

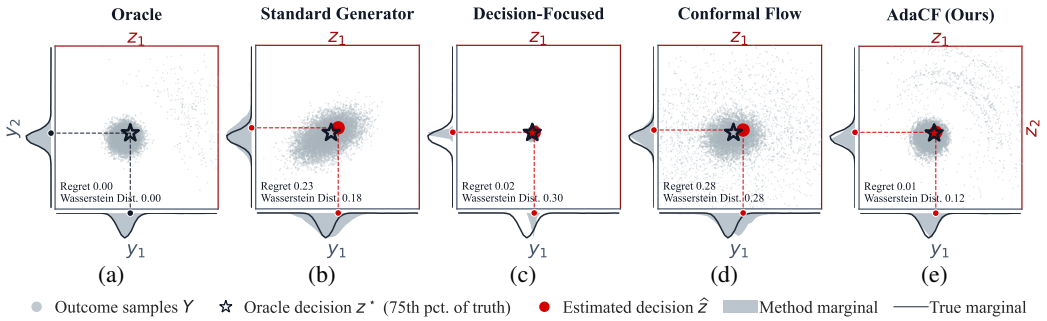


Figure 1: **Comparison of `AdaCF` with other generators** on a two-item newsvendor problem, whose optimal decision is the coordinatewise 75th conditional quantile. Panel (a) is the true conditional distribution at a fixed context and panels (b)–(e) are estimates, with marginals and markers for the oracle and estimated decisions; `AdaCF` best matches both the distribution and the decision.

For example, the chance of an extreme electricity price may depend strongly on current conditions, and a price spike may be more likely in one market than another. A standard conformal flow uses the same distribution over conformal levels at every context and does not learn how the distribution of outcomes changes from one level to another. It can therefore be calibrated while assigning probability in ways that lead to poor decisions.

Our key observation is that a nonconformity score separates this remaining freedom into two parts. The *across-level* distribution describes how likely an outcome is to fall at each score level given the context; the *within-level* distribution describes which outcomes occur at a given level. Both matter for decision making: the first determines when severe outcomes are likely, while the second determines their form. We propose `AdaCF`, an adaptive conformal flow that learns these two distributions from data. Predictive fitting encourages the resulting distribution to represent observed outcomes, and an optional decision loss can direct the fit toward a specified task. A rank-based correction then calibrates the learned level distribution using held-out data. Thus, the learned distribution can adapt to the context and the decision while satisfying a finite-sample *marginal* conformal validity guarantee. The guarantee concerns coverage averaged over contexts and calibration data; it does not assert conditional coverage at every context.

Figure 1 illustrates the distinction on a two-item newsvendor problem. The decision depends on the 75th conditional quantile of each item's demand. A conventionally trained generator produces a plausible predictive distribution but misplaces these decision-relevant quantiles (panel b). A generator trained for the decision recovers the desired quantiles while distorting other parts of the distribution (panel c). A standard conformal flow satisfies its calibration requirement but allocates probability poorly for this context (panel d). By adapting probability across and within score levels, `AdaCF` more closely matches both the true distribution and the oracle decision in this example (panel e).

Our contributions are threefold. First, we formulate decision-focused distribution learning with a conformal calibration constraint and construct `AdaCF` to learn the context-dependent allocation of probability across and within score levels. Second, we establish finite-sample marginal validity for its calibrated sampler and, at the population level, decompose its predictive error into the errors of these two components, characterize the cost of restricting either component, and bound decision regret in terms of predictive error for bounded losses. Third, experiments on synthetic problems and electricity price data show that a single `AdaCF` distribution can improve predictive fidelity and decision quality over standard conformal flows, while remaining competitive with models trained separately for individual decisions.

Related Work. Conformal prediction (CP) provides uncertainty sets with finite-sample, distribution-free coverage under exchangeability (Vovk et al., 2005; Shafer & Vovk, 2008). Split conformal methods enable efficient calibration of modern predictors (Lei et al., 2018; Romano et al., 2019), with extensions addressing distribution shift (Tibshirani et al., 2019; Barber et al., 2023) and general risk control (Bates et al., 2021; Angelopoulos et al., 2024); see Angelopoulos & Bates (2023) for a review. These developments have broadened the scope of CP to time-series forecasting (Gibbs & Candès, 2021; Xu & Xie, 2021), functional outputs and operator models (Lei et al., 2015; Harris & Liu, 2025),

and prediction and inference with noisy or unobserved targets (Einbinder et al., 2024; Zheng et al., 2026). Conformal sets can also be built from conditional cumulative distribution function (CDF) estimates (Chernozhukov et al., 2021), conditional densities (Izbicki et al., 2022), or samples from generative models (Wang et al., 2023; Zheng & Zhu, 2024); complementary work moves beyond sets to full predictive CDFs (Vovk et al., 2019) and to predictive distributions built by optimal transport (Ndiaye, 2025) or score-induced flows (Harris, 2026).

Contextual stochastic optimization selects decisions that minimize expected cost under the conditional distribution of uncertain parameters given observed contexts (Shapiro et al., 2021; Sadana et al., 2025). Predict-then-optimize approaches separate predictive modeling from downstream optimization, using estimated conditional distributions or weighted empirical outcomes to construct decisions (Bertsimas & Kallus, 2020; Kallus & Mao, 2023). Decision-focused learning instead trains predictive models using the downstream loss (Donti et al., 2017; Wilder et al., 2019; Elmachtoub & Grigas, 2022). Recent work extends this approach to generative models that produce outcomes tailored to downstream objectives (Wang et al., 2026; Zhao et al., 2026). The relative merits of these paradigms depend on model specification and problem structure: theoretical comparisons establish statistical advantages for predict-then-optimize under well-specified models, while decision-focused training can improve performance under misspecification (Hu et al., 2022; Elmachtoub et al., 2023). Our approach combines the two paradigms, learning a full predictive distribution for predict-then-optimize with an optional decision-focused term for a specified task, while remaining conformally calibrated.

A growing literature combines conformal calibration with downstream decision making. One direction uses conformal prediction sets as uncertainty sets in robust optimization to obtain probabilistic performance guarantees (Johnstone & Cox, 2021; Sun et al., 2023; Patel et al., 2024; Chen et al., 2026), with extensions to end-to-end uncertainty-set learning (Chenreddy & Delage, 2024; Yeh et al., 2025) and inverse optimization (Lin et al., 2024). Complementary approaches assess or calibrate the risk of candidate decisions (Zhou et al., 2026; Zhou & Zhu, 2026), or derive action policies with safety and risk guarantees (Lekeufack et al., 2024; Kiyani et al., 2025). Conformal predictive distributions have also been used directly for decision making (Vovk & Bendtsen, 2018). Our work builds on this distributional perspective by learning context-dependent probability allocation across and within conformal levels. The resulting predictive distribution preserves marginal calibration with the conformal sets and supports sample-average approximation across downstream objectives, with regret guarantees linking decision quality to the predictive error.

2 Problem Setup

2.1 Conformal Predictive Distributions

Let $(X, Y) \sim P^\star$, with $X \sim P^\star_X$ and $Y \mid X = x \sim P^\star_{Y|X}(\cdot \mid x)$. Given a fixed predictor $\hat{f} : \mathcal{X} \to \mathcal{Y}$ and a nonconformity score $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, split conformal prediction constructs sets satisfying

$$\mathbb{P}\{Y_{\text{test}} \in \mathcal{C}_\alpha(X_{\text{test}})\} \geq 1 - \alpha, \quad \alpha \in (0, 1). \tag{1}$$

Specifically, let $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$, $S_i = S(X_i, Y_i)$, and $\widehat{\tau}_\alpha$ be the empirical $\lceil (n+1)(1-\alpha) \rceil / n$-quantile of the calibration scores. Then $\mathcal{C}_\alpha(x) = \{y \in \mathcal{Y} : S(x, y) \leq \widehat{\tau}_\alpha\}$ satisfies equation 1 when the calibration and test observations are exchangeable. The guarantee is marginal over the calibration data and test pair.

These sets do not specify probabilities for the outcomes they contain. Conformal flows (Harris, 2026) turn the nested sets into a sampleable conformal predictive distribution (CPD). Define the boundary $\partial\mathcal{C}_\alpha(x) = \{y \in \mathcal{Y} : S(x, y) = \widehat{\tau}_\alpha\}$. Under the regularity conditions of Lemma 5 in Appendix A, a hitting map $\Pi_{x,\alpha} : \mathcal{Y} \to \mathcal{Y}$ sends an initial point to this boundary for every $\alpha \in [1/(n+1), 1)$:

$$S\big(x, \Pi_{x,\alpha}(y_0)\big) = \widehat{\tau}_\alpha, \quad \text{for almost every } y_0 \in \mathcal{Y}. \tag{2}$$

For a base distribution $\mu(\cdot \mid x)$, let $\nu(\cdot \mid x, \alpha) = (\Pi_{x,\alpha})_\# \mu(\cdot \mid x)$ denote the resulting distribution on the boundary. Mixing uniformly over levels gives $P^{\text{CPD}}(\cdot \mid x) = \int_0^1 \nu(\cdot \mid x, a)\, da$. Assuming $S(X, Y)$ has a continuous distribution, this CPD satisfies

$$\mathbb{P}_{\widetilde{Y} \sim P^{\text{CPD}}(\cdot|x)} \Big\{ \widetilde{Y} \in \mathcal{C}_\alpha(x) \Big\} \geq 1 - \alpha, \quad \alpha \in (0, 1),$$

with equality on $\mathcal{A}_n := \{k/(n+1) : k = 2, \dots, n\}$; see Appendix A for the behavior between grid points. This statement concerns draws from the CPD at a fixed context. In contrast, equation 1 concerns the true response and holds marginally over contexts; it does not imply $P^\star_{Y|X}(\mathcal{C}_\alpha(x) \mid x) \geq 1 - \alpha$ for every x.

2.2 Decision Making with Calibrated Predictive Distributions

Given a feasible set $\mathcal{Z}$ and a loss $\ell : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$, the oracle decision is

$$z^\star(x) \in \arg\min_{z \in \mathcal{Z}} \mathbb{E}_{P^\star_{Y|X}(\cdot|x)}[\ell(z, Y)].$$

A predictive distribution $Q_x := Q(\cdot \mid X = x)$ instead yields

$$z_Q(x) \in \arg\min_{z \in \mathcal{Z}} \mathbb{E}_{Q_x}[\ell(z, Y)]. \tag{3}$$

We seek a predictive distribution that yields low decision regret while remaining compatible with the nested conformal sets. For $\widetilde{Y} \mid X \sim Q(\cdot \mid X)$, this gives the constrained formulation

$$\begin{aligned} \min_{Q} \quad & \mathbb{E}_{(X,Y)\sim P^\star}[\ell(z_Q(X), Y) - \ell(z^\star(X), Y)] \\ \text{s.t.} \quad & \mathbb{P}\Big\{\widetilde{Y} \in \mathcal{C}_\alpha(X)\Big\} \geq 1 - \alpha, \quad \alpha \in (0,1), \quad \text{with equality for } \alpha \in \mathcal{A}_n. \end{aligned} \tag{4}$$

The coverage probability averages over the calibration data, test context, and generated response. Equality on $\mathcal{A}_n$ excludes distributions that systematically over-cover, such as a point mass at $\hat{f}(x)$. Under exchangeability, $P^\star_{Y|X}$ satisfies the constraint and attains zero regret. Yet neither zero regret for one task nor conformal validity identifies the true conditional distribution. We therefore use predictive fit to distinguish among valid distributions, with an optional decision-focused term when a particular task is specified.

3 Methodology

We propose `AdaCF`, a method that learns predictive distributions along the geometry induced by the nonconformity score. The score decomposes any conditional distribution into two factors: the distribution across score levels, and the distribution of positions within a given level. Conformal calibration does not identify either factor, and standard conformal flows fix both: the level distribution is uniform at every context, and a single base distribution is reused at every level, although neither restriction is required for marginal validity. `AdaCF` instead learns an adaptive level distribution $\pi_\phi(\cdot \mid x)$ and a level-dependent base distribution $\mu_\psi(\cdot \mid x, \alpha)$, and generates a response by drawing a level and an initial point and mapping the point onto the corresponding conformal boundary; keeping the level distribution uniform on average over contexts preserves the conformal guarantee. Section 3.1 formalizes this estimator, Section 3.2 characterizes the role of each factor in predictive error and decision regret, and Figure 2 summarizes the framework and the intuition for its two components.

3.1 Score-Induced Decomposition of Predictive Uncertainty

Across- and within-level uncertainty. To describe how responses are distributed across nonconformity score levels, we measure the nonconformity score on the scale of conformal levels, the index α of the sets $\mathcal{C}_\alpha(x)$. Let $F(s) := \mathbb{P}\{S(X,Y) \leq s\}$ denote the marginal population CDF of the score under $P^\star$, and define the level of a response as

$$A := 1 - F\big(S(X,Y)\big).$$

Intuitively, A indexes the smallest conformal set containing Y: with population thresholds, $Y \in \mathcal{C}_\alpha(X)$ exactly when $\alpha \leq A$. If F is continuous, A is uniformly distributed on $(0,1)$ marginally over contexts and responses, but its conditional distribution given X can vary with the context (Figure 2b). A standard conformal flow, in contrast, draws its level $\widetilde{A}$ from $\mathrm{Unif}(0,1)$ at every context. At context x, define the level law and the within-level response law as

$$\begin{aligned} \pi^\star(\cdot \mid x) &:= \mathrm{Law}(A \mid X = x), \\ P^\star_{Y|X,A}(\cdot \mid x, \alpha) &:= \mathrm{Law}(Y \mid X = x, A = \alpha). \end{aligned}$$

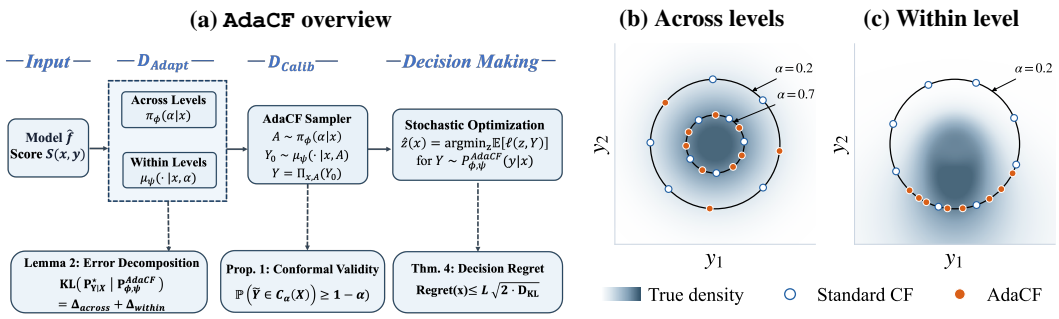


Figure 2: **Overview and learning components of AdaCF.** Panel (a) summarizes the AdaCF pipeline and its theoretical guarantees. Panels (b) and (c) illustrate the intuition for allocation across score levels and within a fixed level, respectively. Shading shows the true conditional density; the markers illustrate standard CF and AdaCF allocations.

Conditioning on A yields the decomposition

$$P^\star_{Y|X}(dy \mid x) = \int_0^1 P^\star_{Y|X,A}(dy \mid x, \alpha)\, \pi^\star(d\alpha \mid x). \tag{5}$$

Intuitively, the level law $\pi^\star$ allocates probability *across* levels, while $P^\star_{Y|X,A}$ allocates position *within* a level. A standard conformal flow sets the first to $\text{Unif}(0,1)$ at every context and pushes the same base distribution to every level, so it recovers equation 5 only when both restrictions hold. The decomposition therefore identifies two learning targets: a context-dependent level law and a context- and level-dependent within-level law.

Adaptive conformal predictive distribution. Motivated by equation 5, AdaCF learns a context-dependent level distribution $\pi_\phi(\cdot \mid x)$ and a context- and level-dependent base distribution $\mu_\psi(\cdot \mid x, \alpha)$, with parameters ϕ and ψ. A calibrated hitting map then pushes the base distribution onto the corresponding boundary, giving the within-level distribution $\nu_\psi(\cdot \mid x, \alpha) := (\Pi_{x,\alpha})_{\#}\mu_\psi(\cdot \mid x, \alpha)$. Given data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{m+n}$, independent of the data used to fit $\hat{f}$, we partition it into an adaptation set $\mathcal{D}_{\text{adapt}}$ of size m and a calibration set $\mathcal{D}_{\text{cal}}$ of size n. We learn π_ϕ and μ_ψ on $\mathcal{D}_{\text{adapt}}$, and use $\mathcal{D}_{\text{cal}}$ to calibrate the hitting map and the learned level distribution, so that the generated responses satisfy the feasibility constraint equation 4.

On $\mathcal{D}_{\text{adapt}}$, given a context x, AdaCF draws $\widetilde{A} \sim \pi_\phi(\cdot \mid x)$ and $\widetilde{Y}_0 \sim \mu_\psi(\cdot \mid x, \widetilde{A})$, and returns $\widetilde{Y} = \Pi_{x,\widetilde{A}}(\widetilde{Y}_0)$. The resulting predictive distribution is

$$P^{\text{AdaCF}}_{\phi,\psi}(B \mid x) = \int_0^1 \nu_\psi(B \mid x, \alpha)\, \pi_\phi(d\alpha \mid x), \qquad B \subseteq \mathcal{Y} \text{ measurable},$$

which mirrors the population decomposition equation 5. To be consistent with the true level law, the learned one should be uniform marginally over contexts, $\mathbb{E}_{X \sim P^\star_X}[\pi_\phi(\cdot \mid X)] = \text{Unif}(0,1)$; under this balance condition, the generated responses satisfy equation 4.

On $\mathcal{D}_{\text{cal}} = \{(X_j, Y_j)\}_{j=1}^n$, we compute the conformal thresholds $\widehat{\tau}_\alpha$ and construct the hitting map as in Harris (2026) (see Appendix A). Because $P^\star_X$ is unknown, the balance condition cannot be enforced exactly during learning, so we also calibrate the level distribution. We draw one reference level $A'_j \sim \pi_\phi(\cdot \mid X_j)$ for each calibration context and, at a test context x, replace the level draw by

$$\widetilde{A} = \frac{\#\{j : A'_j < A'\} + \varepsilon}{n+1}, \qquad A' \sim \pi_\phi(\cdot \mid x), \quad \varepsilon \sim \text{Unif}(0,1). \tag{6}$$

This is a randomized rank correction that rebalances the learned level distribution to be uniform marginally over test contexts, while preserving the context-dependent ordering of levels. The following proposition shows that the calibrated sampler satisfies equation 4 in finite samples.

Proposition 1 (Finite-sample validity of AdaCF)**.** *Suppose π_ϕ and μ_ψ are fitted independently of $\mathcal{D}_{\text{cal}}$, and the test context X_{test} is exchangeable with the calibration contexts. Let $\widetilde{A}$ be given by*

equation 6 at $x = X_{\text{test}}$, $\widetilde{Y}_0 \sim \mu_\psi(\cdot \mid X_{\text{test}}, \widetilde{A})$, *and* $\widetilde{Y} = \Pi_{X_{\text{test}},\widetilde{A}}(\widetilde{Y}_0)$, *where the hitting map is built from the thresholds* $\widehat{\tau}_\alpha$ *on* $\mathcal{D}_{\text{cal}}$. *Then*

$$\mathbb{P}\Big\{\widetilde{Y} \in \mathcal{C}_\alpha(X_{\text{test}})\Big\} \geq 1-\alpha, \quad \alpha \in (0,1), \quad \textit{with equality for } \alpha \in \mathcal{A}_n,$$

where the probability is jointly over $\mathcal{D}_{\text{cal}}$, X_{test}, *and the sampling randomness of* $\widetilde{Y}$.

The proof is given in Appendix A.

Learning the adaptive distribution. The score-induced decomposition suggests estimating the two components separately on $\mathcal{D}_{\text{adapt}} = \{(X_i, Y_i)\}_{i=1}^m$. For each observation, let $s_i = S(X_i, Y_i)$ and define its empirical conformal level as $a_i := (1 + \sum_{j=1}^m \mathbf{1}\{s_j > s_i\})/(m+1)$, with randomized tie-breaking if needed; a_i is the empirical counterpart of A and lies in $(0,1)$. We learn the two components by

$$\begin{aligned}
\widehat{\phi} &\in \arg\min_\phi \underbrace{-\frac{1}{m}\sum_{i=1}^m \log \pi_\phi(a_i \mid X_i)}_{\mathcal{L}_{\text{across}}(\phi)} + \lambda \underbrace{D\Big(\frac{1}{m}\sum_{i=1}^m \pi_\phi(\cdot \mid X_i),\, \text{Unif}(0,1)\Big)}_{\mathcal{L}_{\text{unif}}(\phi)}, \\
\widehat{\psi} &\in \arg\min_\psi \underbrace{-\frac{1}{m}\sum_{i=1}^m \log \mu_\psi(Y_i \mid X_i, a_i)}_{\mathcal{L}_{\text{within}}(\psi)} + \gamma \underbrace{\frac{1}{m}\sum_{i=1}^m \ell\big(z_{Q_\psi}(X_i), Y_i\big)}_{\mathcal{L}_{\text{dec}}(\psi)},
\end{aligned} \tag{7}$$

where D is a discrepancy between distributions on $(0,1)$, such as the Kullback–Leibler divergence, and $\lambda, \gamma \geq 0$. $\mathcal{L}_{\text{across}}$ fits how the severity of the response varies with the context. $\mathcal{L}_{\text{unif}}$ encourages the learned level law to be uniform on average over contexts, as the true one is; exact uniformity is enforced by the calibration step equation 6, so the penalty only keeps that step close to the identity. $\mathcal{L}_{\text{within}}$ lets the kinds of outcomes change with severity, which a base shared across levels cannot represent. $\mathcal{L}_{\text{dec}}$ makes the fit decision-focused, where Q_ψ is the adaptation-stage predictive distribution formed from $\pi_{\widehat{\phi}}$ and μ_ψ, using hitting-map thresholds computed on $\mathcal{D}_{\text{adapt}}$.

3.2 Theoretical Guarantees

We now explain why learning both components matters, in three steps. First, the predictive error of `AdaCF` splits exactly into an across-level and a within-level error (Lemma 2). Second, fixing either component, as standard conformal flows do, leaves an error that no amount of data or model capacity removes (Corollary 3). Third, the two errors together bound the regret of every bounded decision loss, so a distribution accurate in both components makes good decisions for every task (Theorem 4). The analysis concerns the population counterpart of `AdaCF`, in which the empirical thresholds are replaced by population quantiles of the score (Assumption 1), under the regularity conditions of Harris (2026), which let the hitting map reach each boundary while preserving position on it (Lemma 5); only Corollary 3 uses the latter. Finite-sample validity is covered by Proposition 1, and proofs are given in Appendix A.

Lemma 2 (Decomposition of predictive error)**.** *Under Assumption 1, let* $\pi_\phi(\cdot \mid x)$ *be a measurable level kernel and* $\nu_\psi(\cdot \mid x, \alpha)$ *a measurable family of probability measures supported on* $\partial\mathcal{C}_\alpha(x)$. *Then, for* $P_X^\star$*-almost every* x, *the predictive distribution* $P_{\phi,\psi}^{\text{AdaCF}}(\cdot \mid x)$ *satisfies*

$$\begin{aligned}
&D_{\text{KL}}\Big(P_{Y|X}^\star(\cdot \mid x) \,\|\, P_{\phi,\psi}^{\text{AdaCF}}(\cdot \mid x)\Big) \\
&\quad = \underbrace{D_{\text{KL}}(\pi^\star(\cdot \mid x) \,\|\, \pi_\phi(\cdot \mid x))}_{\Delta_{\text{across}}(x)} + \underbrace{\int_0^1 D_{\text{KL}}\Big(P_{Y|X,A}^\star(\cdot \mid x, \alpha) \,\|\, \nu_\psi(\cdot \mid x, \alpha)\Big)\, \pi^\star(d\alpha \mid x)}_{\Delta_{\text{within}}(x)}.
\end{aligned} \tag{8}$$

The decomposition holds for any `AdaCF` distribution, and the model classes need not contain $\pi^\star(\cdot \mid x)$ or $P_{Y|X,A}^\star(\cdot \mid x, \alpha)$. Its two terms correspond to the two components and expose the cost of fixing either one. To quantify this cost for the within-level component, we need a notion of the same

position on different boundaries. Let $U := u_x(Y)$ denote the gradient line through the response, which the hitting map preserves (Lemma 5); for the ℓ_2 score $\|y - \hat{f}(x)\|_2^2$, U is the direction from the prediction. Since the level A fixes the boundary and U the position on it, (A, U) determines Y. The following corollary quantifies the irreducible cost of fixing either component in terms of $I(A; X)$ and $I(U; A \mid X)$, where I denotes (conditional) mutual information under $P^\star$.

Corollary 3 (Information-theoretic cost of restricted adaptation)**.** *Under Assumption 1, suppose the hitting map satisfies Lemma 5 (i)–(iii). Restricting the level distribution to be the same for every context, or the base distribution to be the same at every level, incurs the minimum expected errors*

$$\inf_{\pi_\phi(\cdot|x)\equiv\pi} \mathbb{E}_X\left[\Delta_{\text{across}}(X)\right] = I(A; X), \qquad \inf_{\mu_\psi(\cdot|x,\alpha)\equiv\mu(\cdot|x)} \mathbb{E}_X\left[\Delta_{\text{within}}(X)\right] = I(U; A \mid X).$$

Corollary 3 quantifies the gap between the two constructions: even with unlimited data and model capacity, a standard conformal flow misses the true conditional distribution by at least $I(A; X) + I(U; A \mid X)$ in expected KL divergence, whereas `AdaCF` can close this gap entirely. Each term is a property of the data: $I(A; X)$ measures how strongly the severity of outcomes depends on the context, and $I(U; A \mid X)$ how strongly the kind of outcome depends on its severity. These two terms indicate which adaptation matters for a given problem (Figure 5). To connect these errors to decisions, let $\widehat{z}(x)$ minimize the $P^{\text{AdaCF}}_{\phi,\psi}(\cdot \mid x)$-risk as in equation 3, and define $\text{Reg}_x := \mathbb{E}_{P^\star_{Y|X}(\cdot|x)}[\ell(\widehat{z}(x), Y) - \ell(z^\star(x), Y)]$.

Theorem 4 (Decision regret)**.** *Under the conditions of Lemma 2, suppose* $0 \le \ell(z, y) \le L < \infty$. *Then, for $P^\star_X$-almost every x and every feasible set $\mathcal{Z}$ over which the true and predictive risk minima are attained,*

$$\text{Reg}_x \le L\sqrt{2D_{\text{KL}}\left(P^\star_{Y|X}(\cdot \mid x) \,\|\, P^{\text{AdaCF}}_{\phi,\psi}(\cdot \mid x)\right)} = L\sqrt{2\left[\Delta_{\text{across}}(x) + \Delta_{\text{within}}(x)\right]}. \tag{9}$$

Intuitively, Theorem 4 says that predictive fidelity controls decision regret: whatever the downstream task, the regret of an `AdaCF` decision is at most a fixed multiple of the square root of its predictive error, and the same holds on average over contexts. With $\gamma = 0$, the learning objectives in Section 3.1 aim to reduce these two sources of predictive error, so a single `AdaCF` distribution can serve all downstream tasks; $\gamma > 0$ additionally tailors the within-level fit to a specified task. The proof and a score-based special case are given in Appendix A.

4 Experimental Results

We evaluate `AdaCF` on contextual stochastic optimization in two settings: synthetic data-generating processes (DGPs) with known conditional laws, paired with various decision problems, and electricity price data from the Australian Energy Market Operator (AEMO; Appendix B.2), in which a retailer decides whether to insure against price spikes in neighboring states.

Baselines and metrics. We compare `AdaCF` with **Gen-SAA**, a conditional generator used for sample-average approximation (SAA); **Gen-DFL**, a generator trained through the downstream loss; **Direct-DFL**, which predicts decisions directly; and **Weighted-SAA**, which reweights the empirical responses of similar contexts (Bertsimas & Kallus, 2020). As ablations, we include **CF**, the standard conformal flow, and **CF-Across** and **CF-Within**, which learn only the level law or only the level-dependent base. We measure distributional fidelity by the Wasserstein and energy distances, decision quality by regret relative to the oracle decision, and calibration by the average coverage error of generated responses over a grid of levels; lower is better for all. We set $\gamma = 0$ throughout, so that one fitted distribution serves every decision; Appendix B.6 studies $\gamma > 0$. Details are in Appendix B.

Synthetic benchmark. Figure 3 presents results on a synthetic *surge* DGP with a 10-dimensional context, in which rare heavy-tailed deviations of the response vary in probability, magnitude, and direction with the context. A single `AdaCF` distribution attains the best fidelity among the generative methods and lies on the fidelity–regret frontier (panel a); like the other conformal variants, it stays close to nominal coverage, whereas the non-conformal baselines miss it by up to 19 percentage points (panel b). Panel (c) shows that, unlike CF, `AdaCF` adapts the mass assigned to each score

level across contexts. The gain is consistent across different decisions, where AdaCF outperforms the task-agnostic baselines and remains competitive with task-specific DFL (panel d). Gen-DFL fails on the two reservation decisions, whose indicator losses it can train only through a smoothed surrogate. Table 1 reports both DGPs. On the regime-tail DGP, whose deviations point in nearly the same direction at every severity ($I(U; A \mid X) \approx 0.01$ nats), the level law carries the gain, and CF-Across already matches AdaCF, as Corollary 3 predicts. DGP definitions, decision problems, and tuning are given in Appendix B.1.

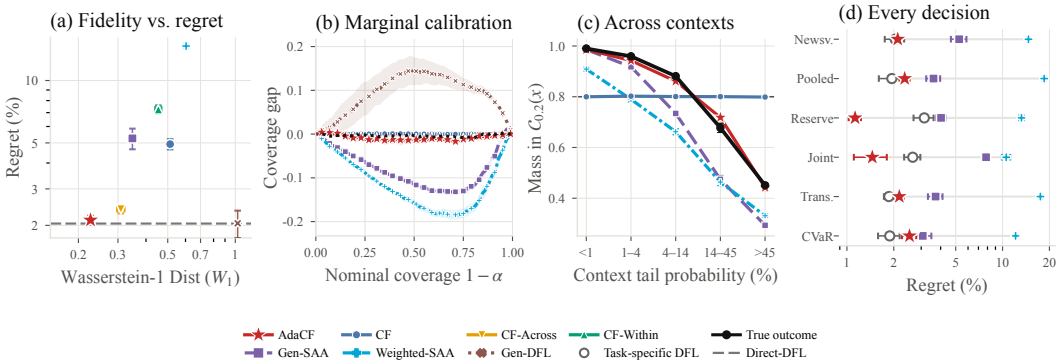


Figure 3: **Synthetic results on the surge DGP. (a)** Newsvendor regret against fidelity. **(b)** Coverage gap $\mathbb{P}\{\widetilde{Y} \in \mathcal{C}_\alpha(X)\} - (1-\alpha)$ of generated responses. **(c)** Mass on the conformal 80% set against the surge probability of the context. **(d)** Regret across different decision problems.

Table 1: **Decision regret and distributional fidelity results**. Regret is in % of the oracle cost for newsvendor (NV), capacity (Cap.), per-item and joint reservation (Res., J. res.), transshipment (Trans.), and CVaR capacity; fidelity is the energy distance and the Wasserstein-1 distance W_1 to the true conditional law. In each column, the two lowest distinct values are in bold; ties are retained. †Retrained separately for every decision.

	Surge								Regime tail							
	Regret (%)						Fidelity		Regret (%)						Fidelity	
Method	NV	Cap.	Res.	J. res.	Trans.	CVaR	Energy	W_1	NV	Cap.	Res.	J. res.	Trans.	CVaR	Energy	W_1
CF	4.9	6.0	2.2	3.2	5.7	12.2	0.149	0.512	3.1	3.7	4.3	3.5	3.4	3.6	0.051	0.213
CF-Across	2.4	2.5	**1.0**	**2.6**	2.3	2.8	**0.078**	**0.309**	**2.4**	3.0	**3.2**	**3.2**	2.7	1.6	**0.038**	0.159
CF-Within	7.3	9.5	4.2	3.8	9.0	13.5	0.127	0.454	3.1	4.1	4.3	3.5	3.7	3.6	0.051	0.222
AdaCF	**2.1**	**2.4**	**1.1**	**1.5**	**2.2**	2.5	**0.052**	**0.226**	2.5	3.6	**3.4**	**3.3**	3.2	1.6	**0.039**	**0.157**
Gen-SAA	5.3	3.6	4.0	7.8	3.7	3.1	0.096	0.348	3.7	2.9	6.0	7.2	3.0	**1.1**	0.048	**0.154**
W-SAA	14.7	18.5	13.2	10.6	17.5	12.1	0.221	0.604	7.8	6.6	12.1	11.9	6.7	6.3	0.131	0.398
Gen-DFL†	**2.0**	2.7	154.0	89.3	**1.9**	**1.9**	6.377	3.946	**2.4**	**2.4**	31.6	28.1	**1.9**	1.7	5.541	3.483
Direct-DFL†	**2.0**	**1.9**	3.1	**2.6**	2.4	**2.1**	–	–	**1.6**	**1.7**	5.2	5.1	**2.1**	**1.3**	–	–

Real-world case study: insuring against electricity price spikes. Given current market features, a retailer decides in advance, for each of South Australia and Victoria, whether to insure the next half-hour against a price spike above \$300/MWh. Insurance costs a fee c, expressed as a fraction of the loss from an uninsured spike; spikes are rare and their probability varies sharply with market conditions (details in Appendix B.2). On a test day with frequent spikes (Figure 4a), CF and Gen-DFL leave most spikes uninsured, Gen-SAA misses one, and AdaCF insures all of them. Over all test days at $c = 0.05$, AdaCF attains the lowest realized cost, 7% below Gen-SAA (panel b). The cost breakdown explains the ranking: Gen-SAA overstates the probability of a spike and pays the highest fees, whereas CF, whose level law is the same in every context, and Gen-DFL leave the most spikes uninsured. The same fitted AdaCF distribution remains cheaper than Gen-SAA at nearly every fee without refitting, while Gen-DFL does not improve on Gen-SAA even at the fee on which it was trained. Appendix B.3 adds a case study that predicts solar power ramps from sky images, where AdaCF lowers the cost of CF by 38–48% and matches Gen-SAA with a smaller calibration error.

Ablation study. We verify Corollary 3 on a controlled DGP in which the two information terms vary separately (Figure 5): the cost of a uniform level law grows with $I(A; X)$, the cost of a base shared

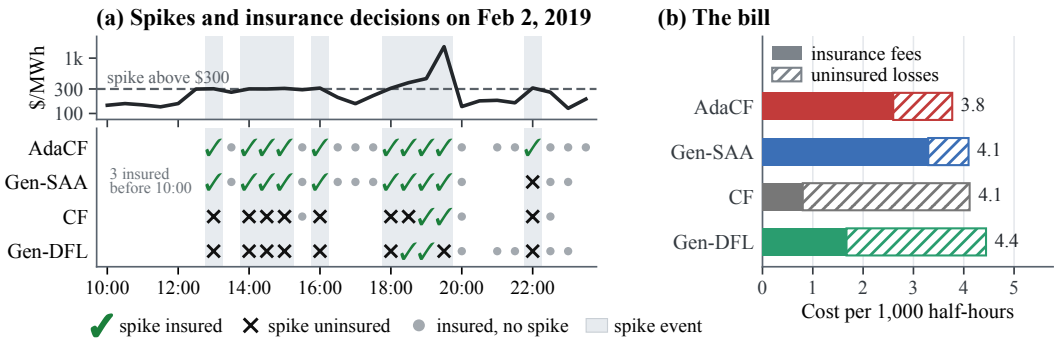


Figure 4: **Price-spike insurance on AEMO data. (a)** Insurance decisions on February 2, 2019 at $c = 0.05$; the line is the higher of the two state prices, and shaded half-hours are spikes. **(b)** Realized cost per 1,000 test half-hours, split into insurance fees and uninsured losses.

across levels grows with $I(U; A \mid X)$, and learning the corresponding component removes most of each cost, underscoring the need to learn both components. In particular, the energy distance of CF grows linearly with $I(A; X)$ (correlation 0.99), while that of `AdaCF` stays flat. Appendices B.4–B.6 further study the effects of the adaptation size, the base generator, and the decision weight γ. The base-generator study shows when the conformal layer helps: `AdaCF` corrects what the base gets wrong about how severity varies with the context, lowering the regret of Gen-SAA by up to 2.8 points with under-fitting Gaussian mixtures and by 17 points with an over-fitted one, whereas with flows, VAEs, or implicit generators that already fit the surge DGP, Gen-SAA is 0.3–0.5 points better, since `AdaCF` fits its base on a subset of the data.

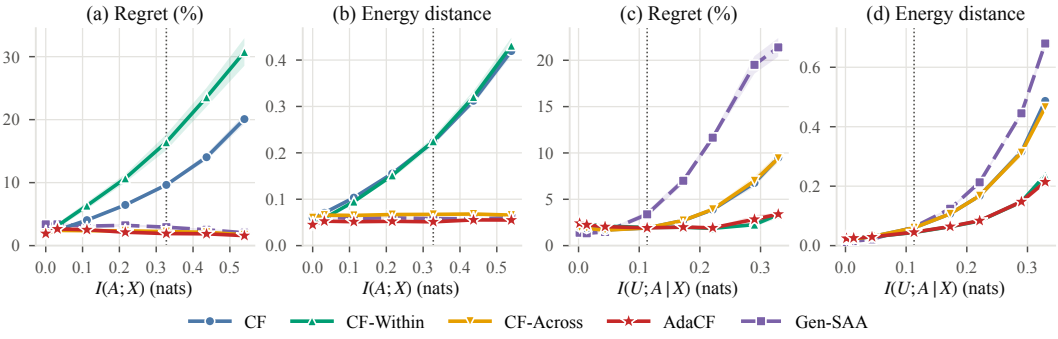


Figure 5: **The cost of each fixed component grows with its information term**. **(a)–(b)** Increasing $I(A; X)$, how strongly severity depends on the context, at $I(U; A \mid X) = 0.11$. **(c)–(d)** Increasing $I(U; A \mid X)$, how strongly the direction of a response depends on its severity, through the surge size at $I(A; X) = 0$. Dotted lines mark the *Both* and *Within only* DGPs of Table 5.

5 Conclusion and Future Work

We studied how to adapt predictive distributions to downstream decisions under a conformal calibration constraint. `AdaCF` learns how probability is allocated across and within conformal levels, retains finite-sample marginal validity, and improves decisions over standard conformal flows; at the population level, the errors of its two components bound regret for every bounded loss. Several directions remain open. The nonconformity score fixes the geometry of the levels, so learning a score whose levels align with the loss, or with context-dependent tail behavior, could make the two components easier to learn. Finite-sample rates would say how much data reliable decisions require, and conditional or localized validity would strengthen the marginal guarantee. Extensions to dependent time series and to unbounded losses remain challenges.

AI USE STATEMENT

In this work, we used generative AI tools to write and polish text and to write experiment code, including data processing, model training, and evaluation scripts. We did not use generative AI tools to fabricate or alter data or results; all reported numbers come from executed code. The authors reviewed all AI-assisted work: code was inspected, results were checked against independent reruns and sanity checks, proofs were verified by hand, and references were checked against the original sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Assumption 1 and Lemma 5 state all conditions used by our theoretical results, and Appendix A gives complete proofs. Appendix B.1 specifies the data-generating processes, decision problems, data splits, and tuning budget for the synthetic benchmarks, and Appendices B.2 and B.3 give the data sources, features, and splits for the two real-data case studies, all of which use publicly available data. Appendix B also records the computing hardware. Anonymized code that reproduces all experiments and figures can be provided upon request and will be released publicly upon acceptance.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Stephen Bates, Anastasios N Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

Andrew Butler and Roy H. Kwon. Integrating prediction in mean-variance portfolio optimization. *Quantitative Finance*, 23(3):429–452, 2023. doi: 10.1080/14697688.2022.2162432.

Shuyi Chen, Ferdinando Fioretto, Feng Qiu, and Shixiang Zhu. Global-decision-focused neural odes for proactive grid resilience management. *IEEE Transactions on Smart Grid*, 2025.

Shuyi Chen, Wenbin Zhou, and Shixiang Zhu. Learning polyhedral conformal sets for robust optimization. *arXiv preprint arXiv:2605.08506*, 2026.

Abhilash Reddy Chenreddy and Erick Delage. End-to-end conditional robust optimization. In *Conference on Uncertainty in Artificial Intelligence*, pp. 736–748, 2024.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

Priya Donti, Brandon Amos, and J Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bat-Sheva Einbinder, Shai Feldman, Stephen Bates, Anastasios N. Angelopoulos, Asaf Gendler, and Yaniv Romano. Label noise robustness of conformal prediction. *Journal of Machine Learning Research*, 25(328):1–66, 2024. URL `https://jmlr.org/papers/v25/23-1549.html`.

Petros Ellinas, Vassilis Kekatos, and Georgios Tsaousoglou. Decision-focused learning under decision dependent uncertainty for power systems with price-responsive demand. *Electric Power Systems Research*, 235:110665, 2024. doi: 10.1016/j.epsr.2024.110665.

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022.

Adam N Elmachtoub, Henry Lam, Haofeng Zhang, and Yunfan Zhao. Estimate-then-optimize versus integrated-estimation-optimization versus sample average approximation: A stochastic dominance perspective. *arXiv preprint arXiv:2304.06833*, 2023.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672, 2021.

Trevor Harris. Flow-based conformal predictive distributions. *arXiv preprint arXiv:2602.07633*, 2026.

Trevor Harris and Yan Liu. Locally adaptive conformal inference for operator models. *arXiv preprint arXiv:2507.20975*, 2025.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Fast rates for contextual linear optimization. *Management Science*, 68(6):4236–4245, 2022.

Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. CD-split and HPD-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32, 2022.

Chancellor Johnstone and Bruce Cox. Conformal uncertainty sets for robust optimization. In *Conformal and Probabilistic Prediction and Applications*, pp. 72–90, 2021.

Nathan Kallus and Xiaojie Mao. Stochastic optimization forests. *Management Science*, 69(4): 1975–1994, 2023.

Shayan Kiyani, George J Pappas, Aaron Roth, and Hamed Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. In *International Conference on Machine Learning*, pp. 30943–30965, 2025.

Jing Lei, Alessandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1–2):29–43, 2015.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Jordan Lekeufack, Anastasios N Angelopoulos, Andrea Bajcsy, Michael I Jordan, and Jitendra Malik. Conformal decision theory: Safe autonomous decisions from imperfect predictions. In *IEEE International Conference on Robotics and Automation*, pp. 11668–11675, 2024.

Bo Lin, Erick Delage, and Timothy C Y Chan. Conformal inverse optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 63534–63564, 2024.

Eugene Ndiaye. Beyond uncertainty sets: Leveraging optimal transport to extend conformal predictive distribution to multivariate settings. *arXiv preprint arXiv:2511.15146*, 2025.

Yash P Patel, Sahana Rayan, and Ambuj Tewari. Conformal contextual robust optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2485–2493. PMLR, 2024.

Matteo Pedone, Raffaele Argiento, and Francesco C. Stingo. Personalized treatment selection via product partition models with covariates. *Biometrics*, 80(1):ujad003, 2024. doi: 10.1093/biomtc/ujad003.

Min Qian and Susan A. Murphy. Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39(2):1180–1210, 2011. doi: 10.1214/10-AOS864.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000. doi: 10.21314/JOR.2000.038.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Utsav Sadana, Abhilash Chenreddy, Erick Delage, Alexandre Forel, Emma Frejinger, and Thibaut Vidal. A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*, 320(2):271–289, 2025.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics, third edition, 2021.

Chunlin Sun, Linyu Liu, and Xiaocheng Li. Predict-then-calibrate: A new perspective of robust contextual LP. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17713–17741, 2023.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel J Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2526–2536, 2019.

Vladimir Vovk and Claus Bendtsen. Conformal predictive decision making. In *Conformal and Probabilistic Prediction and Applications*, pp. 52–62, 2018.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Vladimir Vovk, Jieli Shen, Valery Manokhin, and Min-ge Xie. Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, 108:445–474, 2019.

Prince Wang, Shuyi Chen, Jinhao Liang, Ferdinando Fioretto, and Shixiang Zhu. Gen-DFL: Decision-focused generative learning for robust decision making. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 56198–56223, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5c25c15b5b2fd386ab188a918e54c7d5-Paper-Conference.pdf.

Zhendong Wang, Ruijiang Gao, Mingzhang Yin, Mingyuan Zhou, and David Blei. Probabilistic conformal prediction using conditional random samples. In *International Conference on Artificial Intelligence and Statistics*, pp. 8814–8836, 2023.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569, 2021.

Christopher Yeh, Nicolas Christianson, Alan Wu, Adam Wierman, and Yisong Yue. End-to-end conformal calibration for optimization under uncertainty. *Transactions on Machine Learning Research*, 2025.

Zihao Zhao, Christopher Yeh, Lingkai Kong, and Kai Wang. Diffusion-DFL: Decision-focused diffusion models for stochastic optimization. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 36834–36859, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/3e0b96206965f5f05b0b4550c0e73ff0-Paper-Conference.pdf.

Minxing Zheng and Shixiang Zhu. Generative conformal prediction with vectorized non-conformity scores. *arXiv preprint arXiv:2410.13735*, 2024.

Minxing Zheng, Wenbin Zhou, and Shixiang Zhu. Beyond prediction: Conformal inference for latent distributional parameters. *arXiv preprint arXiv:2608.03607*, 2026. URL https://arxiv.org/abs/2608.03607.

Wenbin Zhou and Shixiang Zhu. Calibrating decision robustness via inverse conformal risk control. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61939.

Wenbin Zhou, Agni Orfanoudaki, and Shixiang Zhu. Conformalized decision risk assessment. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 10581–10610, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/11e0fdc1f9ec9303e5de60670e884e53-Paper-Conference.pdf.

A Assumptions and Proofs

We collect the regularity conditions and proofs for Sections 3.1 and 3.2, in the order in which the results appear. We use $\mathrm{TV}(P,Q) = \sup_B |P(B) - Q(B)|$, and conditional statements are understood for $P_X^\star$-almost every context.

Population regularity.

Assumption 1 (Population regularity)**.** *The context and response spaces are standard Borel, and the score S is measurable. The pooled score CDF $F(s) = \mathbb{P}\{S(X,Y) \le s\}$ is continuous and strictly increasing on its support, and the thresholds are the population quantiles $\tau_\alpha = F^{-1}(1-\alpha)$, so that $\tau_A = S(X,Y)$ almost surely. The conditional laws $\pi^\star(\cdot \mid x)$ and $P^\star_{Y|X,A}(\cdot \mid x, \alpha)$, and all level and base laws considered below, are measurable probability kernels.*

Assumption 1 concerns only the population score distribution: it ensures that distinct levels lie on disjoint boundaries and that the level of a response is determined by the response, which is what makes conditioning on the level exact. Empirical stepwise thresholds do not have this property, so the exact identities below are not asserted for their finite-sample replacements. The following lemma derives the common coordinate representation on an admissible region where the nonconformity flows of Harris (2026) are globally well defined and reach the relevant boundaries.

Properties of the hitting map. Fix a context x and a set of levels under consideration. An *admissible region* is a measurable set $\mathcal{R}_x \subseteq \mathcal{Y}$ that contains Y almost surely under $P^\star_{Y|X}(\cdot \mid x)$ and contains the boundaries $\partial\mathcal{C}_\alpha(x)$ of the levels under consideration; the regions are jointly measurable in (x,y).

Lemma 5 (Properties of the score-gradient hitting map)**.** *Suppose that, for $P_X^\star$-almost every x, the following hold on an admissible region $\mathcal{R}_x$:*

- *(C1) $\nabla_y S(x,\cdot)$ exists, is nonvanishing, and is locally Lipschitz on $\mathcal{R}_x$;*
- *(C2) for every $y_0 \in \mathcal{R}_x$ and every level α under consideration, the score-gradient trajectory started at y_0 exists for all $t \ge 0$ and remains in $\mathcal{R}_x$, together with its limit;*
- *(C3) along each such trajectory, $\|\nabla_y S\|$ is bounded below by a positive constant on its tail.*

Conditions (C2)–(C3) are the hypotheses of Proposition 3.1 of Harris (2026). Then the flow defines a jointly measurable hitting map $\Pi_{x,\alpha}$ on $\mathcal{R}_x$ with the following properties. Write $u_x(y)$ for the gradient line of $S(x,\cdot)$ through y, and $T_{x,\alpha}(u)$ for the point of gradient line u on $\partial\mathcal{C}_\alpha(x)$.

- *(i)* Reaches the boundary: $S(x, \Pi_{x,\alpha}(y_0)) = \tau_\alpha$ *for every $y_0 \in \mathcal{R}_x$, and $\Pi_{x,\alpha}$ fixes every point of $\partial\mathcal{C}_\alpha(x)$.*
- *(ii)* Preserves the gradient line: $u_x(\Pi_{x,\alpha}(y_0)) = u_x(y_0)$ *for every $y_0 \in \mathcal{R}_x$.*
- *(iii)* Gradient lines coordinatize every boundary: *for each level α, $T_{x,\alpha}$ is a measurable bijection from the set $\mathcal{U}_x$ of gradient lines in $\mathcal{R}_x$ onto $\partial\mathcal{C}_\alpha(x)$, with inverse the restriction of u_x, and*

$$\Pi_{x,\alpha}\big(T_{x,\beta}(u)\big) = T_{x,\alpha}(u) \qquad \text{for all levels } \alpha, \beta \text{ and all } u \in \mathcal{U}_x.$$

Proof. Throughout, the flow is $y'(t) = -\lambda\big(S(y(t)) - \tau_\alpha\big)\,\nabla S(y(t))/\|\nabla S(y(t))\|^2$.

(i) This is Proposition 3.1 of Harris (2026). Along the flow, the score error $\varepsilon(t) = S(y(t)) - \tau_\alpha$ satisfies $\varepsilon'(t) = -\lambda\varepsilon(t)$, so (C2) gives $S(y(t)) \to \tau_\alpha$, and (C3) gives convergence to a unique point of $\partial\mathcal{C}_\alpha(x)$. The velocity vanishes on that boundary, so its points are fixed.

(ii) The velocity is parallel to $\nabla_y S$, so a trajectory stays on the gradient line of its initial point; by (i) its limit lies on $\partial\mathcal{C}_\alpha(x)$, hence on that line.

(iii) By (C1), gradient lines are the integral curves of a nonvanishing, locally Lipschitz vector field, hence unique and non-merging, and S is strictly increasing along them. Each gradient

line therefore meets $\partial\mathcal{C}_\alpha(x)$ at most once, and by (i) applied to any of its points, at least once; so $T_{x,\alpha}$ is a bijection with inverse u_x, and the commutation identity follows from (ii). For measurability, fix a reference level, say $1/2$, and encode $u_x(y)$ by $\Pi_{x,1/2}(y)$, so that $T_{x,\alpha}$ is the restriction of $\Pi_{x,\alpha}$ to the reference boundary; the finite-time ODE solution maps are jointly measurable and their pointwise limits define the hitting maps, so the coordinate maps and their inverses are jointly measurable.

□

Examples. For the ℓ_2 score, gradient lines are rays from $\hat{f}(x)$ and u_x is the direction. For a one-dimensional absolute-error score with positive thresholds, $\mathcal{U}_x = \{-1, 1\}$ and $T_{x,\alpha}(u) = \hat{f}(x) + u\tau_\alpha$, so u_x is the sign of the residual. Property (iii) is what makes "the same position on different boundaries" meaningful: u_x identifies position on each boundary, rather than merely recording a feature that the flow happens to preserve.

On the conditions. (C2) cannot be dropped: a smooth score with nonvanishing gradient may have gradient lines that never reach a given boundary. For $S(y) = \exp(e^{y_1} \sin y_2)$, the line $y_2 = \pi/2$ has $S > 1$ throughout and never meets $\{S = 1/2\}$, and the flow leaves every compact set in finite time. For the ℓ_2 score $S(x, y) = \|y - \hat{f}(x)\|_2^2$ used in our experiments, (C1)–(C3) hold on $\mathcal{R}_x = \mathcal{Y} \setminus \{\hat{f}(x)\}$ for every level with $\tau_\alpha > 0$: the gradient $\nabla_y S = 2(y - \hat{f}(x))$ is smooth and vanishes only at $\hat{f}(x)$; each trajectory stays on the ray from $\hat{f}(x)$ through its initial point, along which S moves monotonically from $S(x, y_0)$ to τ_α, so it never reaches $\hat{f}(x)$ and exists for all $t \geq 0$; and $\|\nabla_y S\| = 2S^{1/2} \geq 2\min\{S(x, y_0), \tau_\alpha\}^{1/2} > 0$ along it. This region contains Y almost surely whenever $P^\star_{Y|X}(\cdot \mid x)$ has no atom at $\hat{f}(x)$.

Finite-sample validity.

Proof of Proposition 1. Write $R \coloneqq \#\{j : A'_j < A'\}$ for the rank in equation 6. Since π_ϕ is fitted independently of $\mathcal{D}_{\text{cal}}$ and of X_{test}, and each reference level is drawn from $\pi_\phi(\cdot \mid \cdot)$ at its own context, the pairs $(X_1, A'_1), \ldots, (X_n, A'_n), (X_{\text{test}}, A')$ are exchangeable. Hence R is uniform on $\{0, \ldots, n\}$, ties having probability zero when $\pi_\phi(\cdot \mid x)$ is continuous (and being broken at random otherwise), and $\widetilde{A} = (R + \varepsilon)/(n + 1)$ with $\varepsilon \sim \text{Unif}(0, 1)$ independent is exactly uniform on $(0, 1)$. In particular $\mathbb{P}\{\widetilde{A} \geq \alpha\} = 1 - \alpha$ for every α.

Fix $\alpha \geq 1/(n + 1)$. On the event $\{\widetilde{A} \geq \alpha\}$, the hitting-map identity equation 2 with the empirical thresholds gives $S(X_{\text{test}}, \widetilde{Y}) = \widehat{\tau}_{\widetilde{A}} \leq \widehat{\tau}_\alpha$, because $\widehat{\tau}_\beta$ is nonincreasing in β by construction, so

$$\{\widetilde{A} \geq \alpha\} \subseteq \{\widetilde{Y} \in \mathcal{C}_\alpha(X_{\text{test}})\}.$$

This inclusion holds pointwise for any realized thresholds, so their dependence on $\mathcal{D}_{\text{cal}}$ does not affect the argument, and the base distribution enters only through the position of $\widetilde{Y}$ on its boundary, which does not affect membership in $\mathcal{C}_\alpha(X_{\text{test}})$. Taking probabilities gives $\mathbb{P}\{\widetilde{Y} \in \mathcal{C}_\alpha(X_{\text{test}})\} \geq \mathbb{P}\{\widetilde{A} \geq \alpha\} = 1 - \alpha$. For $\alpha < 1/(n + 1)$, $\widehat{\tau}_\alpha = +\infty$ and $\mathcal{C}_\alpha(X_{\text{test}}) = \mathcal{Y}$, so the bound is trivial. Draws with $\widetilde{A} < 1/(n + 1)$, which occur with probability $1/(n + 1)$, target an empty boundary, since $\widehat{\tau}_{\widetilde{A}} = +\infty$, and are mapped to the boundary of the largest calibration score; the lower bound holds for any such convention, since these draws lie outside $\{\widetilde{A} \geq \alpha\}$ for every $\alpha \geq 1/(n + 1)$, but equality uses this one. Let $\alpha = k/(n + 1)$ with $2 \leq k \leq n$. With distinct calibration scores, a draw with $1/(n + 1) \leq \widetilde{A} < \alpha$ lies on a boundary with $\widehat{\tau}_{\widetilde{A}} > \widehat{\tau}_\alpha$, and a draw with $\widetilde{A} < 1/(n + 1)$ lies on the boundary $\widehat{\tau}_{1/(n+1)} > \widehat{\tau}_\alpha$; neither is in $\mathcal{C}_\alpha(X_{\text{test}})$, so the inclusion above is an equality and the bound is attained. For $k = 1$, the draws with $\widetilde{A} < 1/(n + 1)$ lie in $\mathcal{C}_{1/(n+1)}(X_{\text{test}})$, so the coverage there is 1. Only the independence of π_ϕ from $\mathcal{D}_{\text{cal}}$ is used; since $\mathcal{L}_{\text{dec}}$ in equation 7 trains only ψ, the guarantee holds for every $\gamma \geq 0$. □

Decomposition of predictive error. KL values may be infinite; the identity is understood in the extended sense.

Proof of Lemma 2. Fix x and define $\mathrm{lev}_x(y) := 1 - F(S(x, y))$. Under the true law, $A = \mathrm{lev}_x(Y)$ almost surely by Assumption 1; under the generated law, $\widetilde{A} = \mathrm{lev}_x(\widetilde{Y})$ almost surely, because $\nu_\psi(\cdot \mid x, \alpha)$ is supported on $\partial\mathcal{C}_\alpha(x)$, on which $S(x, \cdot) = \tau_\alpha$. Consequently,

$$(\mathrm{lev}_x)_\# P^\star_{Y|X}(\cdot \mid x) = \pi^\star(\cdot \mid x), \qquad (\mathrm{lev}_x)_\# P^{\mathrm{AdaCF}}_{\phi,\psi}(\cdot \mid x) = \pi_\phi(\cdot \mid x).$$

Augmenting a response by its deterministic level leaves KL unchanged, since the augmentation has the measurable inverse $(\alpha, y) \mapsto y$ on its image. Disintegrating the two augmented laws gives

$$\pi^\star(d\alpha \mid x) P^\star_{Y|X,A}(dy \mid x, \alpha), \qquad \pi_\phi(d\alpha \mid x)\nu_\psi(dy \mid x, \alpha).$$

The chain rule for relative entropy therefore yields equation 8. If $\pi^\star(\cdot \mid x)$ is not absolutely continuous with respect to $\pi_\phi(\cdot \mid x)$, both the marginal KL and the joint KL are infinite; otherwise the conditional form of the chain rule applies, again with extended values. □

The two learning objectives of Section 3.1 target the two conditional laws in this identity. At the population level, the excess conditional log loss of the level objective equals $\mathbb{E}_X \Delta_{\mathrm{across}}(X)$ whenever the log-loss risks are well defined. The within-level objective is a conditional maximum-likelihood fit of $\mu_\psi(\cdot \mid x, \alpha)$ at the observed response, which lies on the boundary of its level; its population minimizer over a well-specified class is the true within-level law, but identifying its excess risk with Δ_{within} requires densities relative to a reference measure on the boundary, which an ambient density does not by itself provide. With $\gamma > 0$, the true within-level law remains a population minimizer when the class is well specified and the level law is correct, since the true distribution attains the smallest expected decision loss.

Information cost of restricted adaptation. Corollary 3 uses properties (i)–(iii) of Lemma 5. Let $U \coloneqq u_X(Y)$ under $P^\star$, and let base measures be supported on the admissible region of that lemma. The model classes in Corollary 3 range over all measurable level kernels and all such base kernels, subject only to the restrictions stated there; in particular, any measurable coordinate law λ_x can be induced by a shared base, for example $(T_{x,1/2})_\#\lambda_x$, and any law $\lambda_{x,\alpha}$ by a level-dependent base.

Proof of Corollary 3. Let $\rho_{x,\alpha} := \mathrm{Law}(U \mid X = x, A = \alpha)$ and $\rho_x := \mathrm{Law}(U \mid X = x)$. A base $\mu(\cdot \mid x)$ shared across levels induces a single coordinate law $\lambda_x = (u_x)_\#\mu(\cdot \mid x)$ at every level. Lemma 5(ii)–(iii) gives

$$P^\star_{Y|X,A}(\cdot \mid x, \alpha) = (T_{x,\alpha})_\#\rho_{x,\alpha}, \qquad (\Pi_{x,\alpha})_\#\mu(\cdot \mid x) = (T_{x,\alpha})_\#\lambda_x.$$

KL is invariant under these measurable bijections. Applying the chain rule to the joint law of (A, U) conditional on $X = x$ yields

$$\begin{aligned}\int_0^1 D_{\mathrm{KL}}\big(P^\star_{Y|X,A}(\cdot \mid x, \alpha) \,\|\, (\Pi_{x,\alpha})_\#\mu(\cdot \mid x)\big)\, \pi^\star(d\alpha \mid x) &= \int_0^1 D_{\mathrm{KL}}(\rho_{x,\alpha}\|\lambda_x)\, \pi^\star(d\alpha \mid x) \\ &= I(U; A \mid X = x) + D_{\mathrm{KL}}(\rho_x\|\lambda_x).\end{aligned}$$

Thus the best shared coordinate law is $\lambda_x = \rho_x$, with expected within-level error $I(U; A \mid X)$, and a level-dependent base realizes $\lambda_{x,\alpha} = \rho_{x,\alpha}$ and makes this error zero.

For a context-independent level law π, the marginal uniformity of A under Assumption 1 gives $\mathbb{E}_X D_{\mathrm{KL}}(\pi^\star(\cdot \mid X)\|\pi) = I(A; X) + D_{\mathrm{KL}}(\mathrm{Unif}(0, 1)\|\pi)$, minimized at $\pi = \mathrm{Unif}(0, 1)$ with value $I(A; X)$, while an adaptive level law realizes $\pi_\phi(\cdot \mid x) = \pi^\star(\cdot \mid x)$ and makes this error zero. The level and coordinate laws can be chosen independently, and the class description above supplies base measures for each minimizing coordinate law. Adding the two minima through equation 8 gives the minimum for the standard class in Corollary 3; adapting both components makes both terms vanish. The characterization of when the standard class contains the truth follows from the equivalence between vanishing mutual information and the corresponding conditional independence. □

These minima concern the unrestricted classes described above. A finite parametrization or an imperfect optimizer may incur additional approximation or estimation error, and a positive KL minimum alone does not establish positive regret for any particular downstream task.

Decision regret. For a decision z, write $R_\star(z \mid x) := \mathbb{E}_{P^\star_{Y|X}(\cdot|x)}\ell(z, Y)$ and $R_{\mathrm{AdaCF}}(z \mid x) := \mathbb{E}_{P^{\mathrm{AdaCF}}_{\phi,\psi}(\cdot|x)}\ell(z, Y)$.

Proof of Theorem 4. For any measurable g with values in $[0, L]$ and probability measures ρ, ρ',

$$\left|\int g\, d(\rho - \rho')\right| \le L\, \mathrm{TV}(\rho, \rho').$$

Optimality of $\widehat{z}(x)$ under $P^{\mathrm{AdaCF}}_{\phi,\psi}(\cdot \mid x)$ implies

$$\begin{aligned}\mathrm{Reg}_x &\le [R_\star - R_{\mathrm{AdaCF}}](\widehat{z}(x) \mid x) + [R_{\mathrm{AdaCF}} - R_\star](z^\star(x) \mid x)\\ &\le 2 \sup_{z \in \mathcal{Z}} |R_{\mathrm{AdaCF}}(z \mid x) - R_\star(z \mid x)|.\end{aligned}$$

Applying the preceding inequality to the loss gives $\mathrm{Reg}_x \le 2L\, \mathrm{TV}(P^\star_{Y|X}(\cdot \mid x), P^{\mathrm{AdaCF}}_{\phi,\psi}(\cdot \mid x))$, and Pinsker's inequality together with equation 8 yields equation 9. The expected form follows from Jensen's inequality applied to the concave map $t \mapsto \sqrt{t}$. □

If $\ell(z, y) = c(z, S(x, y))$, then on each boundary both conditional expected losses equal $c(z, \tau_\alpha)$; hence, holding the level law fixed, changing the within-level laws does not affect any decision risk. For unbounded losses, the regret bound requires additional tail or moment control.

B Experimental Details and Additional Results

This section gives the implementation details for the synthetic benchmark (Appendix B.1) and the two real-data case studies (Appendices B.2 and B.3), followed by three ablations. The ablations ask, in turn, whether each adaptation repairs the failure the theory attributes to it (Appendix B.4), when the conformal layer improves on a given base generator (Appendix B.5), and what the optional decision term adds when a single decision is known (Appendix B.6). Unless otherwise stated, `AdaCF` uses $\gamma = 0$: one predictive distribution is fitted per data split and evaluated on every downstream decision without task-specific retraining.

Hardware. Experiments were run on a desktop with an Intel Core Ultra 7 265F CPU and an NVIDIA GeForce RTX 5070 Ti GPU (16 GB VRAM), and on a MacBook Pro with an Apple M2 Pro chip (12-core CPU, 19-core GPU, and 32 GB unified memory).

B.1 Synthetic benchmark

Data-generating processes. Both DGPs draw a context $X \sim \mathrm{Unif}([-1, 1]^{10})$ and a two-item response $Y \in \mathbb{R}^2$, read as the demands of two items, and depend on the context only through a few dense linear indices of X with fixed unit vectors.

- **Surge.** With indices $s_k(X) = \sqrt{5}\, u_k^\top X$ for $k = 1, 2, 3$, the response is

$$Y = \begin{cases} m(X) + M\, D(X) & \text{with probability } p(X) = \sigma\big(-2.5 + 2\, s_1(X)\big),\\ m(X) + \omega(X)\, \varepsilon & \text{otherwise,}\end{cases}$$

  where σ is the logistic function, $\varepsilon \sim \mathcal{N}(0, I_2)$, $\omega(X) = 0.6\,\{1 + 0.25 \tanh s_2(X)\}$, and the location is $m(X) = \big(6 + 0.8 \tanh s_1 + 0.1 \tanh s_2,\ 6 + 0.8 \tanh s_1 + 0.9 \tanh s_2\big)$. A surge has a heavy-tailed size $M = 1 + E$ with E exponential of mean 2.5, and a noisy direction $D(X) = |d(X) + 0.35\, \xi|/\|d(X) + 0.35\, \xi\|$ with $\xi \sim \mathcal{N}(0, I_2)$, where $d(X)$ interpolates between two fixed directions in the positive orthant as $s_3(X)$ varies. The surge probability ranges from below 1% to above 45% across contexts, so both the frequency and the form of severe responses depend on the context.
- **Regime tail.** With bounded indices $h_k(X) = \tanh(\sqrt{3}\, v_k^\top X)$ for $k = 1, \ldots, 4$, the response is the bivariate Student-t variable

$$Y = m(X) + \sigma(X)\, Z/\sqrt{W},$$

with $m(X) = (6 + h_1 + h_2,\ 6 + h_1 - h_2)$ and $\sigma(X) = 0.8 \exp\{0.25\, h_4(X)\}$, where $Z \sim \mathcal{N}(0, \Sigma)$ has unit variances and correlation 0.3, and $W \sim \chi^2_{\nu(X)}/\nu(X)$ is shared by the two items, so that extreme deviations hit them together. The degrees of freedom $\nu(X)$ interpolate log-linearly in $h_3(X)$ between 2.2 (heavy tails) and 30 (nearly Gaussian). The direction of a deviation barely depends on its severity: $I(U; A \mid X) \approx 0.01$ nats against $I(A; X) \approx 0.05$.

Decision problems. Table 2 lists the six decisions. The quantile decisions use the critical fractile $q = 0.8$, so a unit of shortage costs $b = q/(1-q) = 4$ times a unit of surplus; the reservation decisions use the demand threshold 8 and the reservation cost 0.15. For a predictive distribution, the decision solves the sample-average approximation of equation 3 with 300 generated responses per context. Regret is reported in percent of the oracle cost, $100\,(\bar{\ell} - \bar{\ell}^\star)/\bar{\ell}^\star$, where $\bar{\ell}$ and $\bar{\ell}^\star$ average the expected loss of the method's decision and of the oracle decision over the test contexts.

Table 2: **Decision problems of the synthetic benchmark.** Here $(t)_+ = \max(t, 0)$, $b = 4$, $\mathrm{Sh} = \sum_i (y_i - z_i)_+$ and $\mathrm{Su} = \sum_i (z_i - y_i)_+$ are the total shortage and surplus, and binary decisions $r_i = 1$ reserve item i.

Decision	Variable	Loss $\ell(z, y)$
Newsvendor	$z \in \mathbb{R}^2$	$\sum_{i=1}^2 \big[b\,(y_i - z_i)_+ + (z_i - y_i)_+\big]$
Capacity	$z \in \mathbb{R}$	$b\,(y_1 + y_2 - z)_+ + (z - y_1 - y_2)_+$
Reservation	$r \in \{0,1\}^2$	$\sum_{i=1}^2 \big[0.15\, r_i + (1 - r_i)\,\mathbf{1}\{y_i > 8\}\big]$
Joint reservation	$r \in \{0,1\}^2$	$0.15 \sum_{i=1}^2 r_i + \mathbf{1}\{\exists\, i :\ y_i > 8,\ r_i = 0\}$
Transshipment	$z \in \mathbb{R}^2$	$b\,\mathrm{Sh} + \mathrm{Su} - b \min(\mathrm{Sh}, \mathrm{Su})$ (unit transfer cost)
CVaR capacity	$(z, \eta) \in \mathbb{R}^2 \times \mathbb{R}$	$0.4 \cdot \frac{1}{2} \sum_i z_i + \eta + \big(\max_i (y_i - z_i)_+ - \eta\big)_+ / (1 - 0.9)$

Implementation and configuration. Table 3 summarizes the configuration. In the synthetic experiments, we use a conditional Gaussian whose mean and diagonal covariance are produced by a multilayer perceptron with two hidden layers of width 64 as the base generative model, so that the methods differ only in how they use the data and remain fair comparisons. The conformal family fits the base generator on the training split, which also defines the ℓ_2 score, centered at the mean of 32 draws of the base generator at each context; it learns the level law and the level-dependent base on the adaptation split and calibrates on the calibration split. The baselines are fitted on the union of the three splits. The level law π_ϕ is a softmax classifier over 20 equal-width level bins, with $\lambda = 0$ in equation 7, since the rank correction equation 6 enforces uniformity exactly. The level-dependent base adds the level as an input to the shared base and is warm-started from it. Gen-DFL uses the same generator architecture trained through a differentiable sample-average approximation with 64 generated responses; Direct-DFL is a multilayer perceptron from the context to the decision, with a sigmoid output for the reservation decisions. Every method receives the same tuning budget of six candidates, selected on a shared validation split by the energy score for distributional methods and by the validation task loss for the DFL models.

B.2 AEMO ELECTRICITY PRICES

Task. Before each half-hour, a retailer decides for each of South Australia and Victoria whether to buy insurance against a price spike, in the form of a cap contract that pays out the part of the price above \$300/MWh. Insurance costs a fee c, expressed as a share of the unit loss incurred when a spiking state is left uninsured, so a predictive distribution Q_x yields the decision

$$r(x) \in \operatorname*{arg\,min}_{r \in \{0,1\}^2}\ c\,(r_1 + r_2) + \mathbb{P}_{Q_x}\big\{\exists\, i :\ P_i > 300,\ r_i = 0\big\},$$

where P_1, P_2 are the two prices. The main text reports $c = 0.05$, the fee on which Gen-DFL is trained.

Data. We use the 30-minute regional price and demand data published by AEMO[1] for South Australia (SA1) and Victoria (VIC1) from January 2019 to December 2020, which give 35,040

[1] https://www.aemo.com.au/energy-systems/electricity/national-electricity-market-nem/data-nem/aggregated-data

Table 3: **Configuration of the synthetic benchmark.**

Data	
Conformal family	3,000 training, 4,000 adaptation, and 1,000 calibration draws
Baselines	all 8,000 draws, pooled
Model selection	1,000 validation draws, shared by all methods
Evaluation	1,000 test contexts; oracle decisions and costs from 2,000–3,000 true draws per context
Repetitions	10 seeds, each with fresh data and initialization
Tuning grids (six candidates each)	
Gen-SAA, conformal base	learning rate $\{10^{-3}, 3 \cdot 10^{-3}\}\times$ steps $\{1{,}500, 3{,}000, 6{,}000\}$
Shared and level-dependent bases	penalty toward the base $\{0, 10^{-3}, 10^{-2}\}\times$ steps $\{1{,}500, 3{,}000\}$
Weighted-SAA	bandwidth $\{0.1, 0.3, 1\} \times \sqrt{d/2} \times$ neighbors $\{30, 300\}$
Gen-DFL	learning rate $\{10^{-3}, 3 \cdot 10^{-3}\}\times$ steps $\{1{,}000, 2{,}000, 4{,}000\}$, per decision
Direct-DFL	learning rate $\{3 \cdot 10^{-3}, 10^{-2}\}\times$ steps $\{800, 1{,}600, 3{,}200\}$, per decision
Metrics	
Regret	sample-average decisions from 300 generated responses per context
Energy distance, W_1	200 generated and 200 true draws per context
Coverage	levels $\alpha \in \{0.05, 0.10, \ldots, 0.95\}$

half-hours; gaps of up to two half-hours are interpolated. The response is the pair of prices on the scale $\operatorname{arcsinh}(P/100)$, which is monotone, so the spike threshold maps to $\operatorname{arcsinh}(3)$. The 14 context features are the time of day, weekday, and month (as sine and cosine), a weekend indicator, the demand in each state, the price of each state in the previous half-hour and at the same half-hour of the previous day, and the highest South Australian price over the previous 24 hours; realized demand stands in for the operator's demand forecast. Spikes occur in 0.69% of half-hours in South Australia and 0.45% in Victoria.

Protocol. Whole days are assigned at random to training (30%), adaptation (30%), calibration (10%), validation (10%), and test (20%) sets, and results are averaged over 10 random assignments. As in Appendix B.1, the conformal family uses the three fitting sets separately, the baselines are fitted on their union, and every method has the same tuning budget; Gen-DFL is trained through a smoothed version of the decision above. The day in Figure 4(a) is the test day of the first assignment with the most half-hours in which both states spiked.

Results. At $c = 0.05$, the cost reduction of `AdaCF` relative to Gen-SAA has a paired standard error of 2.6 percentage points. Across 28 fees from 0.015 to 0.15, the same fitted `AdaCF` distribution is cheaper than Gen-SAA at 26 of them; at the two smallest fees, where insuring almost every half-hour is optimal, Gen-SAA is cheaper. Gen-DFL assigns a spike probability of exactly zero to 97.5% of the test half-hours, which contain 83 of the 237 test spikes; this is the source of its uninsured losses.

B.3 Solar power ramps from sky images

The AEMO study relies on tabular market features. Here the signal is visual: a passing cloud can cut the output of a solar array by a third within minutes, and whether such a ramp is likely can be seen in the sky. The level law must therefore learn from image features how severe a drop is likely to be. The study also adds a second type of decision, a battery reserve, alongside the alert of Appendix B.2.

Task. We predict the drop in photovoltaic (PV) output over the next 10 and 20 minutes, $\mathrm{PV}(t) - \mathrm{PV}(t+h)$, and consider two decisions fixed before any result: a *ramp alert*, the joint alert of Appendix B.2 with a drop of 8 kW (about a quarter of capacity) as the threshold for each horizon and fee $c = 0.05$; and a *battery reserve* for each horizon at critical fractile 0.95, the newsvendor decision in which each kW of shortfall costs 19 times as much as each kW of unused reserve.

Data. We use the SKIPP'D dataset[2] of 64×64 fisheye sky images and one-minute PV output of a roughly 30 kW array at Stanford University from 2017 to 2019, taken every five minutes, which gives 21,070 rows on 453 days. The 42 context features are the current PV output and its values 1, 5, 10, and 15 minutes earlier, its standard deviation over the previous 15 minutes, the time of day and day of year (as sine and cosine), and 32 principal components of an ImageNet-pretrained ResNet-18 embedding of the current sky image. Drops above 8 kW occur in 1.4% of rows over 10 minutes and 1.8% over 20 minutes; over 20 minutes their rate rises from 0.14% in the calmest quintile of the recent variability to 7.7% in the most variable.

Protocol. As in Appendix B.2, with whole days assigned to the five sets. Gen-DFL is trained separately for each decision, and Direct-DFL is omitted.

Results. Table 4 and Figure 6 summarize the results.

- `AdaCF` lowers the cost of CF by 38% for the ramp alert and by 48% for the battery reserve in every assignment, and its energy score by 17%.
- `AdaCF` and Gen-SAA are indistinguishable: their costs are within 3% with paired standard errors of about 4 percentage points, they have the same energy score, and `AdaCF` has the smaller calibration error.
- Each Gen-DFL model is cheaper on the decision it was trained for, by 23% and 11%, but its predictive distribution is far from the data, with an energy score of 8.0 against 0.12, and at an alert fee of 0.2 `AdaCF` is 15% cheaper.
- CF predicts a probability between 1.3% and 2.9% for a drop above 8 kW in every regime, while the observed rate rises from 0.2% to 7.3%; `AdaCF` and Gen-SAA follow the observed rate. Under broken clouds (Figure 6), the 95% reserve of CF is almost zero, while the observed drops at similar times have a 95% quantile of 13.4 kW.

Table 4: **Solar power ramps from sky images** (mean over ten splits; standard errors in parentheses). The alert cost is the realized cost of the ramp alert at $c = 0.05$ per 1,000 test rows; the reserve cost is the realized newsvendor loss of the 95% battery reserve in kW, summed over the two horizons; energy is the energy score of the two PV drops; calibration error is the average gap, over a grid of levels, between the generated mass in the conformal sets and the nominal level; the regime gap is the average, over the five quintiles of the recent variability, of the gap between the generated and the observed mass in the conformal 80% set. †Retrained separately for each decision; its energy score, calibration error, and regime gap are averaged over the two models.

Method	Alert cost	Reserve cost	Energy	Calibration error	Regime gap
CF	21.7 (2.0)	13.37 (0.76)	0.145	0.001	0.205
CF-Across	12.8 (1.1)	6.86 (0.38)	0.120	0.033	0.029
CF-Within	21.6 (2.0)	13.59 (0.72)	0.144	0.001	0.205
`AdaCF`	13.2 (1.2)	6.94 (0.41)	0.120	0.033	0.029
Gen-SAA	13.0 (1.6)	7.00 (0.68)	0.121	0.093	0.027
Weighted-SAA	14.4 (0.7)	7.78 (0.39)	0.125	0.062	0.033
Gen-DFL†	10.2 (0.8)	6.16 (0.34)	8.0	0.328	0.406

B.4 Controlled ablation of across- and within-level dependence

On the benchmarks of Section 4, the severity and the kind of outcome depend on the context at the same time, so the gains cannot be attributed to either component alone. Corollary 3 makes a sharper prediction: a uniform level law costs at least $I(A; X)$, and a base shared across levels costs at least $I(U; A \mid X)$. Here we test it on a DGP in which the two information terms can be switched on separately, so that each adaptation should repair one failure and leave the other untouched.

Design. We draw

$$Y = m(X) + s(X)\, W, \qquad s(X) = 0.8\, e^{\rho\, h_3(X)}, \qquad W = Z + JM\, e - \mathbb{E}[JM]\, e,$$

[2] `https://huggingface.co/datasets/solarbench/SKIPPD`

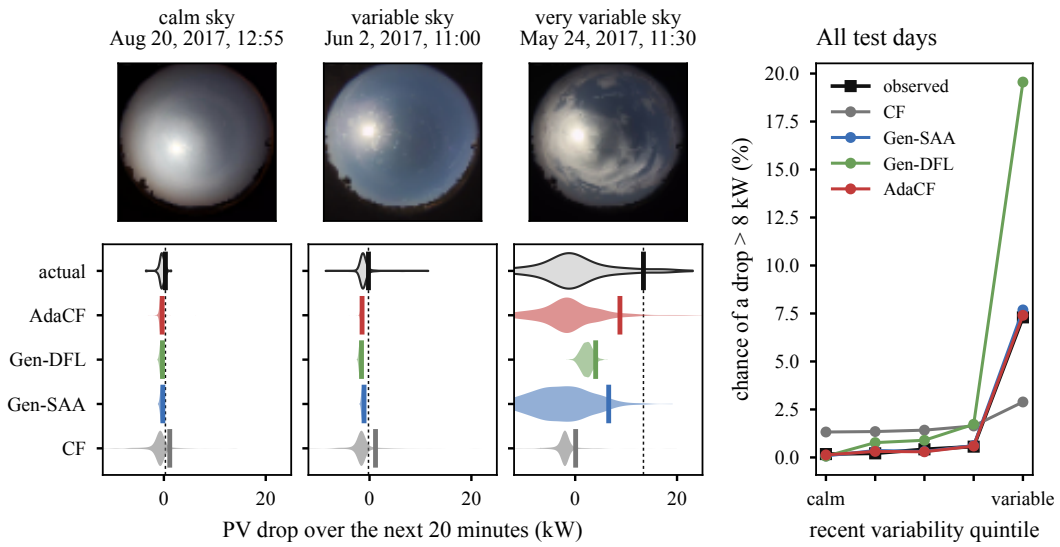


Figure 6: **Solar power ramps from sky images. Left three panels:** sky images at three test times of the first split and, below each, the predicted distributions of the 20-minute PV drop at that time, with each method's 95% battery reserve (tick); the top row (actual) shows the observed drops at all test times of the same variability quintile between 11:00 and 13:00 (73, 203, and 228 times), with their 95% quantile (dotted). **Right:** observed share of 20-minute drops above 8 kW and each method's mean predicted probability, by quintile of the recent variability, over all test days of ten splits.

where m and h_3 are as in the regime-tail DGP, $Z \sim \mathcal{N}(0, I_2)$, the surge indicator J is Bernoulli with probability 0.2κ, the surge size is $M = 1 + E$ with E exponential of mean 2, and $e = (1,1)/\sqrt{2}$. Since W has mean zero, the ℓ_2 score, centered at the mean of the fitted base generator, is approximately centered at $\mathbb{E}[Y \mid X]$. The two parameters switch the two information terms on separately:

- ρ controls $I(A;X)$: with $\rho = 0$ the law of $Y - m(X)$ is the same at every context, so $I(A;X) = 0$; with $\rho = 1$ the spread varies about sevenfold across contexts.
- κ controls $I(U;A \mid X)$: with $\kappa = 0$ the residual is isotropic Gaussian, so the direction is independent of the level; with $\kappa = 1$ the most severe responses are joint surges of both items.

The four DGPs with $\rho, \kappa \in \{0,1\}$ form a 2×2 design in the two information terms, which we compute numerically (Table 5), and the four conformal-flow variants form the 2×2 of the two adaptations. By Corollary 3, the smallest expected KL error each variant can reach is $I(A;X) + I(U;A \mid X)$ for CF, $I(U;A \mid X)$ for CF-Across, $I(A;X)$ for CF-Within, and zero for `AdaCF`. The protocol is that of Appendix B.1.

Results. Each adaptation reduces one failure and only one (Table 5, Figures 7 and 8).

- *Level law.* The severity error of the variants with a uniform level law, CF and CF-Within, is about 0.19 on the two DGPs with $I(A;X) > 0$ and at most 0.045 on the other two; learning the level law brings it to 0.05–0.06.
- *Level-dependent base.* The direction error of the variants with a shared base, CF and CF-Across, is 0.24 on the two DGPs with $I(U;A \mid X) > 0$ and at most 0.05 on the other two. The level-dependent base reduces it to 0.07 and 0.12, but not to its floor, because a Gaussian base concentrates the most severe draws toward the surge direction less sharply than the true law does.
- *Both.* On the DGP with both dependencies, only `AdaCF` corrects both failures, and it attains the lowest energy distance and regret, below CF-Across on all five decisions. The within-level correction pays off only together with the level law: with uniform levels, CF-Within

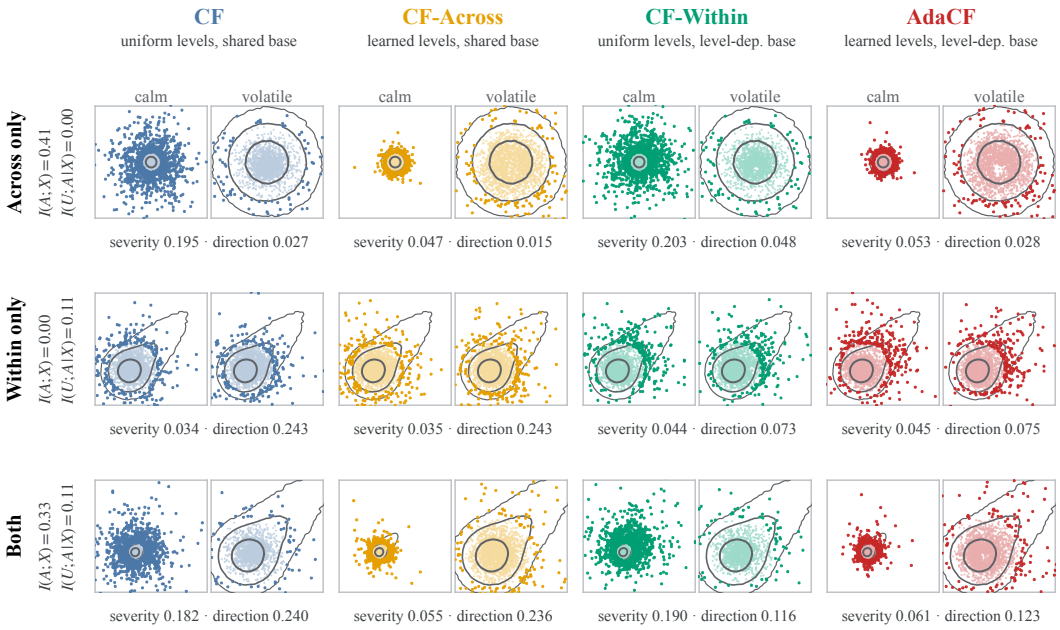


Figure 7: **Controlled ablation.** Samples from each variant (columns) on the across-only, within-only, and combined DGPs (rows), at calm and volatile contexts. Gray contours enclose 50%, 90%, and 99% of the true law; darker dots mark the true top 10% by severity. Numbers report the errors in Table 5.

Table 5: **Controlled ablation of the two adaptations** (mean over ten seeds). The four DGPs form a 2×2 design in the two information terms of Corollary 3 (in nats), and the four variants form the 2×2 of the two adaptations (CF-A: CF-Across; CF-W: CF-Within). The severity error measures the across-level failure and the direction error the within-level failure; their sampling floors are 0.022 and 0.004. Regret is averaged over the five decisions of Appendix B.1. Standard errors are at most 0.003 for the severity error, 0.013 for the direction error, 0.009 for the energy distance, and 1.3 points for regret.

			Severity error				Direction error			
DGP	$I(A;X)$	$I(U;A \mid X)$	CF	CF-A	CF-W	`AdaCF`	CF	CF-A	CF-W	`AdaCF`
Null	0	0	0.030	0.031	0.035	0.036	0.018	0.018	0.029	0.028
Across only	0.41	0	0.195	0.047	0.203	0.053	0.027	0.015	0.048	0.028
Within only	0	0.11	0.034	0.035	0.044	0.045	0.243	0.243	0.073	0.075
Both	0.33	0.11	0.182	0.055	0.190	0.061	0.240	0.236	0.116	0.123

			Energy distance				Regret (%)			
DGP	$I(A;X)$	$I(U;A \mid X)$	CF	CF-A	CF-W	`AdaCF`	CF	CF-A	CF-W	`AdaCF`
Null	0	0	0.018	0.018	0.023	0.022	2.0	1.9	2.5	2.5
Across only	0.41	0	0.156	0.020	0.156	0.023	16.7	2.0	16.1	2.0
Within only	0	0.11	0.060	0.060	0.044	0.045	1.9	1.9	2.0	1.9
Both	0.33	0.11	0.223	0.067	0.225	0.051	9.6	2.2	16.4	1.9

draws severe levels at calm contexts and, with them, surges, so its regret is 16.4% against 9.6% for CF.

- *Null cost.* On the within-only DGP the four variants have the same mean regret, although the level-dependent base improves the direction error and the energy distance; on the null DGP it costs about 0.6 points of regret.

Information and sample-size sweeps. Figure 5 varies one information term at a time.

- $I(A;X)$. With $\kappa = 1$ and $\rho \in \{0, 0.25, \ldots, 1.5\}$, $I(A;X)$ ranges from 0 to 0.54 nats at $I(U;A \mid X) = 0.11$. The energy distance of CF grows linearly, $0.039 + 0.645\, I(A;X)$ (correlation 0.99), while CF-Across and `AdaCF` stay flat.

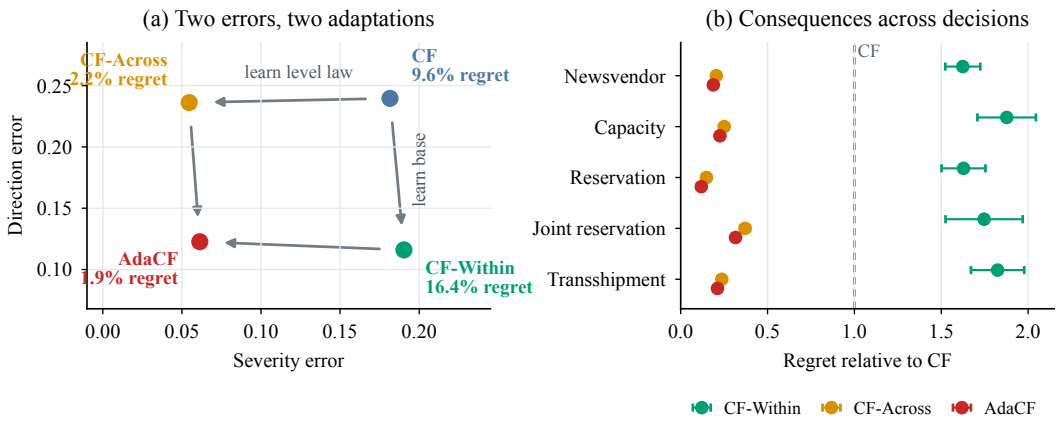


Figure 8: **Interaction of the two adaptations on the combined DGP. (a)** Arrows add a learned level law (horizontal) or a level-dependent base (vertical); labels give mean regret across five decisions and ten seeds. Lower severity and direction error is better. **(b)** Regret on each decision relative to CF, averaged over ten paired seeds; error bars are standard errors across seeds. Learning the within-level law alone lowers direction error but worsens regret when the context-dependent level law is still missing.

- $I(U; A \mid X)$. With $\rho = 0$, so that $I(A; X) = 0$, and the mean of E in $\{0, 0.5, 1, 2, 3, 4, 6, 8\}$, $I(U; A \mid X)$ ranges from 0.002 to 0.33 nats. The level-dependent base lowers the regret of CF-Across by 2.0, 4.2, and 6.0 points at the three largest values, in every seed, and raises it by about 0.4 points below 0.05 nats.
- *Adaptation size.* Figure 9 varies m on the DGP with both dependencies, with $3{,}000$ training and $1{,}000$ calibration draws throughout and Gen-SAA fitted on all $4{,}000 + m$ draws. CF and CF-Within level off at an energy distance of about 0.21, while CF-Across and `AdaCF` decrease to 0.05 and, from $m = 2{,}000$, have lower regret than Gen-SAA in every seed. With the same tuning grid at every m, every conformal variant is worse than Gen-SAA below $m = 1{,}000$, where even the most conservative candidate overfits the adaptation set; at $m \geq 8{,}000$, `AdaCF` has 0.2–0.4 points higher regret than CF-Across despite its lower energy distance.

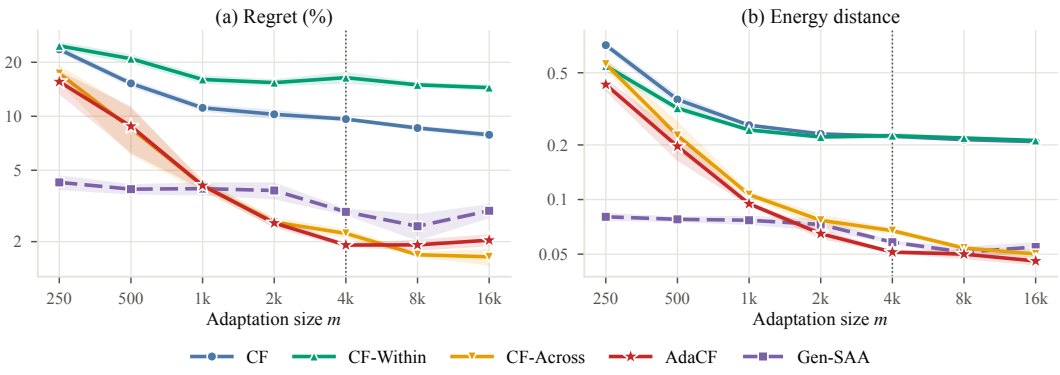


Figure 9: **Effect of the adaptation size** on the DGP with both dependencies (mean $\pm$ s.e. over ten seeds). The variants with a uniform level law level off, while CF-Across and `AdaCF` keep improving; the dotted line marks $m = 4{,}000$, used everywhere else.

B.5 Choice of base generator

`AdaCF` is a layer on top of a base generator: it relearns how probability is spread across score levels and calibrates it on separate data. Its benefit should therefore depend on what the base gets wrong. A

base that already captures how severity varies with the context leaves little to correct, whereas an under-fitting or over-fitted base leaves much. Here we isolate this effect by varying the base family while keeping the data, splits, and tuning protocol fixed, and comparing Gen-SAA with `AdaCF` built on the same base.

Setup. Table 6 and Figure 10 use the surge DGP, with ten seeds and otherwise the protocol of Appendix B.1. All networks have two hidden layers of width 64, and the level-dependent base of `AdaCF` uses the same family as the shared one. The bases are:

- *under-fitting*: conditional Gaussian mixtures with $K = 1, 2, 4$, and 8 components;
- *flexible*: a conditional normalizing flow with six affine coupling layers, a conditional variational autoencoder with a four-dimensional latent variable, and an implicit generator trained by the energy score;
- *reduced data* (1,000 training, 2,000 adaptation, and 500 calibration draws): a flow tuned as usual, and a 16-component mixture over-fitted by 20,000 training steps without model selection.

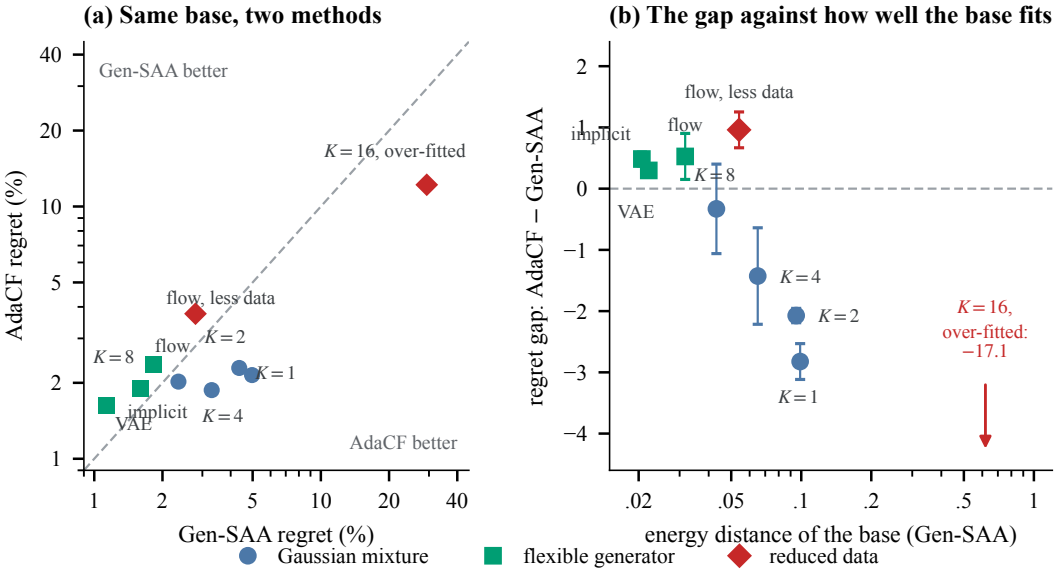


Figure 10: **Choice of base generator on the surge DGP. (a)** Mean regret of Gen-SAA and of `AdaCF` built on the same base generator, over the five decisions of Appendix B.1 and ten seeds; points below the diagonal are bases that `AdaCF` improves. **(b)** The same comparison as a difference, against the energy distance of the base itself, with standard errors over seeds; the over-fitted mixture is off the scale at -17.1. The conformal layer pays off where the base is inaccurate and costs a fraction of a point where it is not.

Results. `AdaCF` is most useful when the base misrepresents how severity varies with the context, and costs a fraction of a point when the base is already accurate.

- With an under-fitting mixture, `AdaCF` lowers the mean regret of Gen-SAA by 2.8 points with one component, 1.4 with four, and an insignificant 0.3 with eight.
- The flexible families fit the surge law closely (energy distance 0.02–0.03), and Gen-SAA then has 0.3–0.5 points lower regret: the conformal layer has little to correct, while it fits its base on a subset of the data.
- With the over-fitted mixture, whose spread is badly miscalibrated, the held-out adaptation and calibration of `AdaCF` cut the regret of Gen-SAA from 29.3% to 12.2%.

Table 6: **Choice of base generator on the surge DGP** (mean over ten seeds). Each row compares Gen-SAA with `AdaCF` built on the same base generator. Regret is averaged over the five decisions of Appendix B.1 (% of the oracle cost), Δ is the paired difference `AdaCF` minus Gen-SAA with its standard error, and the last two columns are energy distances to the true conditional law.

		Regret (%)			Energy distance	
	Base generator	Gen-SAA	`AdaCF`	Δ	Gen-SAA	`AdaCF`
Under-fitting	Mixture, $K = 1$	4.97	2.15	-2.82 ± 0.29	0.099	0.053
	Mixture, $K = 2$	4.36	2.29	-2.07 ± 0.12	0.095	0.061
	Mixture, $K = 4$	3.30	1.87	-1.43 ± 0.79	0.065	0.048
	Mixture, $K = 8$	2.35	2.02	-0.33 ± 0.73	0.043	0.050
Flexible	Flow	1.83	2.35	$+0.53 \pm 0.38$	0.032	0.055
	VAE	1.59	1.89	$+0.30 \pm 0.10$	0.022	0.042
	Implicit (energy score)	1.14	1.62	$+0.48 \pm 0.11$	0.021	0.037
Reduced data	Flow, tuned	2.80	3.76	$+0.96 \pm 0.29$	0.054	0.097
	Mixture, $K = 16$, over-fitted	29.34	12.19	-17.14 ± 1.48	0.619	0.395

B.6 Effect of the decision weight

The main experiments set $\gamma = 0$, so that one fitted distribution serves every decision. When a single decision is known in advance, however, the decision term in equation 7 can steer the within-level fit toward the part of the distribution that this decision uses, much as decision-focused learning does. Because the term trains only μ_ψ, the conformal guarantee is unaffected (Proposition 1); what can change is the fit itself. Here we sweep γ to measure this trade-off: how much the steering helps the trained decision, and what it costs in fidelity and in the regret of decisions it was not trained for.

Setup. We sweep $\gamma \in \{10^{-4}, 10^{-3}, \ldots, 10^{2}\}$ in equation 7, training $\mathcal{L}_{\mathrm{dec}}$ on the newsvendor decision only and evaluating three untrained decisions of Appendix B.1: capacity, joint reservation, and CVaR capacity.

- *Pairing.* Everything upstream of μ_ψ (data, score, level law, thresholds, and reference levels) is fitted once per seed and shared by all values of γ, so the arms differ only in the objective of μ_ψ.
- *Decision term.* $\mathcal{L}_{\mathrm{dec}}$ is evaluated on the adaptation-stage distribution Q_ψ with 64 generated responses per context and a differentiable SAA solve, rescaled by a constant fixed at initialization so that γ is comparable across DGPs.
- *Placebo.* A placebo arm keeps $\gamma = 1$ but permutes the responses in $\mathcal{L}_{\mathrm{dec}}$ within each batch, which preserves the magnitude of the term and removes its signal.
- *DGPs.* The surge DGP of the main text and four DGPs with a two-dimensional context: the surge DGP, a heavy-tail DGP whose tail frequency, reach, and direction vary with the context, a bimodal DGP with context-dependent mode weights and orientation, and a polar bimodal DGP in which severity and mode depend jointly on the context.

The sweep uses ten seeds and the default configuration of our implementation, with a four-component mixture base and no per-method tuning, so its values at $\gamma = 0$ differ from Table 1.

Table 7: **Effect of the decision weight** γ (mean over ten seeds). The decision term is trained on the newsvendor task only; regret (% of the oracle cost) is reported for that task and averaged over the three untrained tasks. The placebo arm uses $\gamma = 1$ with the responses in $\mathcal{L}_{\mathrm{dec}}$ randomly permuted. Calibration error does not depend on γ.

	Trained regret				Untrained regret			Energy			Calib.
DGP	γ=0	1	10	placebo	0	1	10	0	1	10	error
Surge (d=10)	2.6	2.5	3.5	6.0	3.1	3.3	4.4	0.064	0.067	0.101	0.015
Surge (d=2)	1.2	1.0	1.0	3.6	2.2	2.2	2.5	0.027	0.027	0.035	0.013
Heavy tail	1.3	1.1	1.2	4.3	2.2	2.1	2.3	0.092	0.110	0.160	0.012
Bimodal	3.2	3.1	4.0	4.0	3.7	3.8	4.8	0.070	0.076	0.193	0.010
Polar bimodal	2.8	1.6	1.4	4.0	1.6	1.2	1.2	0.031	0.036	0.083	0.011

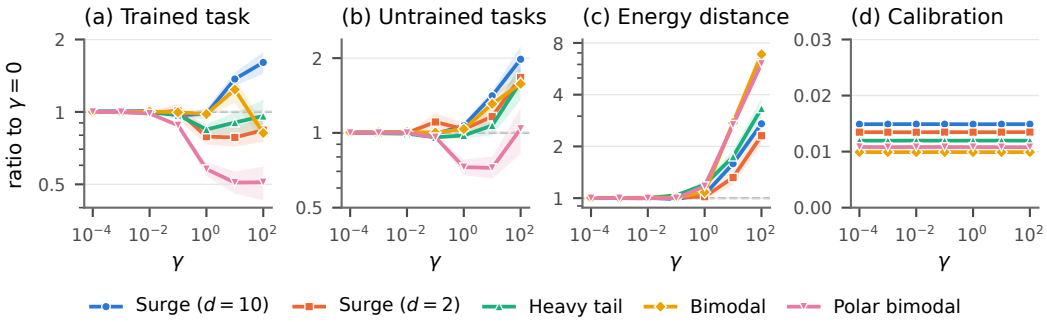


Figure 11: **Effect of the decision weight** γ (ten seeds). **(a)–(c)** Regret on the trained newsvendor task, mean regret on the three untrained tasks, and energy distance to the true conditional law, each as a ratio to its value at $\gamma = 0$ (dashed line), with 68% paired bootstrap bands. **(d)** Calibration error of the generated responses.

Results. Figure 11 and Table 7 show three regularities.

- *Calibration does not depend on* γ, as Proposition 1 predicts: the arms of a seed differ by at most 1.3×10^{-4} in calibration error, the Monte Carlo error of the evaluation.
- *Fidelity degrades as* γ *grows*: the energy distance is nearly unchanged up to $\gamma = 0.1$, rises by at most 20% at $\gamma = 1$, and by up to 2.8-fold at $\gamma = 10$.
- *The decision term helps only where the within-level model is restrictive.* At $\gamma = 1$, it lowers the trained-task regret by 42% on the polar bimodal DGP (paired $t = -6.8$ across seeds), where it also lowers the untrained-task regret by 27%, and by 15–21% on the heavy-tail and two-dimensional surge DGPs. On the main-text surge DGP and the bimodal DGP it brings no significant gain, and at $\gamma = 10$ it raises both regrets on these two DGPs. The placebo is worse than $\gamma = 0$ on every DGP, so the gains come from the decision signal.

A moderate decision weight is therefore worth trying when a single decision is known and the within-level model is restrictive, while $\gamma = 0$ remains the choice for a distribution reused across decisions.