

---

# ECCCoS from the Black Box: Letting Models speak for Themselves

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 Counterfactual Explanations offer an intuitive and straightforward way to explain  
2 Black Box Models but they are not unique. To address the need for plausible  
3 explanations, existing work has primarily relied on surrogate models to learn how  
4 the input data is distributed. This effectively reallocates the task of learning realistic  
5 representations of the data from the model itself to the surrogate. Consequently, the  
6 generated explanations may look plausible to humans but not necessarily conform  
7 with the behaviour of the Black Box Model. We formalise this notion of model  
8 conformity through the introduction of tailored evaluation measures and propose  
9 a novel algorithmic framework for generating **Energy-Constrained Conformal**  
10 **Counterfactuals** that are only as plausible as the model permits. To do so, **ECCCo**  
11 leverages recent advances in energy-based modelling and predictive uncertainty  
12 quantification through conformal inference. Through illustrative examples and  
13 extensive empirical studies, we demonstrate that ECCCoS reconcile the need for  
14 plausibility and model conformity.

## 15 1 Introduction

16 Counterfactual Explanations provide a powerful, flexible and intuitive way to not only explain Black  
17 Box Models but also enable affected individuals to challenge them through the means of Algorithmic  
18 Recourse. Instead of opening the black box, Counterfactual Explanations work under the premise  
19 of strategically perturbing model inputs to understand model behaviour [29]. Intuitively speaking,  
20 we generate explanations in this context by asking simple what-if questions of the following nature:  
21 ‘Our credit risk model currently predicts that this individual’s credit profile is too risky to offer them a  
22 loan. What if they reduced their monthly expenditures by 10%? Will our model then predict that the  
23 individual is credit-worthy?’

24 This is typically implemented by defining a target outcome  $\mathbf{y}^* \in \mathcal{Y}$  for some individual  $\mathbf{x} \in \mathcal{X} = \mathbb{R}^D$   
25 described by  $D$  attributes, for which the model  $M_\theta : \mathcal{X} \mapsto \mathcal{Y}$  initially predicts a different outcome:  
26  $M_\theta(\mathbf{x}) \neq \mathbf{y}^*$ . Counterfactuals are then searched by minimizing a loss function that compares the  
27 predicted model output to the target outcome:  $y_{\text{loss}}(M_\theta(\mathbf{x}), \mathbf{y}^*)$ . Since Counterfactual Explanations  
28 (CE) work directly with the Black Box Model, valid counterfactuals always have full local fidelity by  
29 construction [17]. Fidelity is defined as the degree to which explanations approximate the predictions  
30 of the Black Box Model. This is arguably one of the most important evaluation metrics for model  
31 explanations, since any explanation that explains a prediction not actually made by the model is  
32 useless [16].

33 In situations where full fidelity is a requirement, CE therefore offers a more appropriate solution to  
34 Explainable Artificial Intelligence (XAI) than other popular approaches like LIME [22] and SHAP  
35 [12], which involve local surrogate models. But even full fidelity is not a sufficient condition for  
36 ensuring that an explanation adequately describes the behaviour of a model. That is because two

37 very distinct explanations can both lead to the same model prediction, especially when dealing with  
38 heavily parameterized models:

39           [. . .] deep neural networks are typically very underspecified by the available  
40           data, and [. . .] parameters [therefore] correspond to a diverse variety of compelling  
41           explanations for the data. — Wilson [30]

42 When people talk about Black Box Models, this is usually the type of model they have in mind.

43 In the context of CE, the idea that no two explanations are the same arises almost naturally. Even  
44 the baseline approach proposed by Wachter et al. [29] can yield a diverse set of explanations  
45 if counterfactuals are initialised randomly. This multiplicity of explanations has not only been  
46 acknowledged in the literature but positively embraced: since individuals seeking Algorithmic  
47 Recourse (AR) have unique preferences, Mothilal et al. [17], for example, have prescribed *diversity*  
48 as an explicit goal for counterfactuals. More generally, the literature on CE and AR has brought  
49 forward a myriad of desiderata for explanations, which we will discuss in more detail in the following  
50 section.

## 51 2 Background and Related Work

52 In this section, we provide some background on Counterfactual Explanations and our motivation for  
53 this work. To start off, we briefly introduce the methodology underlying most state-of-the-art (SOTA)  
54 counterfactual generators.

### 55 2.1 Gradient-Based Counterfactual Search

56 While Counterfactual Explanations can be generated for arbitrary regression models [24], existing  
57 work has primarily focused on classification problems. Let  $\mathcal{Y} = (0, 1)^K$  denote the one-hot-encoded  
58 output domain with  $K$  classes. Then most SOTA counterfactual generators rely on gradient descent  
59 to optimize different flavours of the following counterfactual search objective:

$$\mathbf{Z}' = \arg \min_{\mathbf{Z}' \in \mathcal{Z}^M} \{ \text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^*) + \lambda \text{cost}(f(\mathbf{Z}')) \} \quad (1)$$

60 Here  $\text{yloss}$  denotes the primary loss function already introduced above and  $\text{cost}$  is either a single  
61 penalty or a collection of penalties that are used to impose constraints through regularization. Fol-  
62 lowing the convention in Altmeyer et al. [2] we use  $\mathbf{Z}' = \{\mathbf{z}_m\}_M$  to denote the  $M$ -dimensional  
63 array of counterfactual states. This is to explicitly account for the fact that we can generate multiple  
64 counterfactuals  $M$ , as with DiCE [17], and may choose to traverse a latent encoding  $\mathcal{Z}$  of the feature  
65 space  $\mathcal{X}$  where we denote  $f^{-1} : \mathcal{X} \mapsto \mathcal{Z}$ . Encodings may involve simple feature transformations or  
66 more advanced techniques involving generative models, as we will discuss further below.

67 Solutions to Equation 1 are considered valid as soon as the predicted label matches the target label. A  
68 stripped-down counterfactual explanation is therefore little different from an adversarial example.  
69 In Figure 1, for example, we have the baseline approach proposed in Wachter et al. [29] to MNIST  
70 data (centre panel). This approach solves Equation 1 through gradient-descent in the feature space  
71 with a penalty for the distance between the factual  $\mathbf{x}$  and the counterfactual  $\mathbf{x}'$ . The underlying  
72 classifier  $M_\theta$  is a simple Multi-Layer Perceptron (MLP) with good test accuracy. For the generated  
73 counterfactual  $\mathbf{x}'$  the model predicts the target label with high confidence (centre panel in Figure 1).  
74 The explanation is valid by definition, even though it looks a lot like an Adversarial Example [6].  
75 Schut et al. [23] make the connection between Adversarial Examples and Counterfactual Explanations  
76 explicit and propose using a Jacobian-Based Saliency Map Attack (JSMA) to solve Equation 1. They  
77 demonstrate that this approach yields realistic and sparse counterfactuals for Bayesian, adversarially  
78 robust classifiers. Applying their approach to our simple MNIST classifier does not yield a realistic  
79 counterfactual but this one, too, is valid (right panel in Figure 1).

### 80 2.2 From Adversial Examples to Plausible Explanations

81 The crucial difference between Adversarial Examples (AE) and Counterfactual Explanations is one  
82 of intent. While an AE is intended to go unnoticed, a CE should have certain desirable properties.

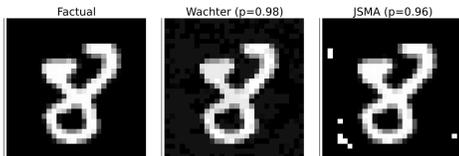


Figure 1: You may not like it, but this is what stripped-down counterfactuals look like. Counterfactuals for turning an 8 (eight) into a 3 (three): original image (left); counterfactual produced using Wachter et al. [29] (centre); and a counterfactual produced using JSMA-based approach introduced by [23].

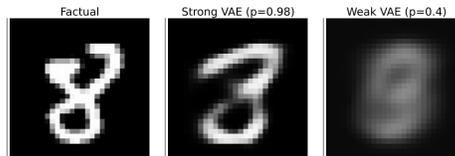


Figure 2: Using surrogates can improve plausibility, but also increases vulnerability. Counterfactuals for turning an 8 (eight) into a 3 (three): original image (left); counterfactual produced using REVISE [9] with a well-specified surrogate (centre); and a counterfactual produced using REVISE [9] with a poorly specified surrogate (right).

83 The literature has made this explicit by introducing various so-called *desiderata*. To properly serve  
 84 both AI practitioners and individuals affected by AI decision-making systems, counterfactuals should  
 85 be sparse, proximate [29], actionable [27], diverse [17], plausible [9, 21, 23], robust [26, 20, 2] and  
 86 causal [11] among other things.

87 Researchers have come up with various ways to meet these desiderata, which have been extensively  
 88 surveyed and evaluated in various studies [28, 10, 19, 4, 8]. Perhaps unsurprisingly, the different  
 89 desiderata are often positively correlated. For example, Artelt et al. [4] find that plausibility typically  
 90 also leads to improved robustness. Similarly, plausibility has also been connected to causality in the  
 91 sense that plausible counterfactuals respect causal relationships [13].

### 92 2.2.1 Plausibility through Surrogates

93 Arguably, the plausibility of counterfactuals has been among the primary concerns and some have  
 94 focused explicitly on this goal. Joshi et al. [9], for example, were among the first to suggest that  
 95 instead of searching counterfactuals in the feature space  $\mathcal{X}$ , we can instead traverse a latent embedding  
 96  $\mathcal{Z}$  that implicitly codifies the data generating process (DGP) of  $\mathbf{x} \sim \mathcal{X}$ . To learn the latent embedding,  
 97 they introduce a surrogate model. In particular, they propose to use the latent embedding of a  
 98 Variational Autoencoder (VAE) trained to generate samples  $\mathbf{x}^* \leftarrow \mathcal{G}(\mathbf{z})$  where  $\mathcal{G}$  denotes the decoder  
 99 part of the VAE. Provided the surrogate model is well-trained, their proposed approach —REVISE—  
 100 can yield compelling counterfactual explanations like the one in the centre panel of Figure 2.

101 Others have proposed similar approaches. Dombrowski et al. [5] traverse the base space of a  
 102 normalizing flow to solve Equation 1, essentially relying on a different surrogate model for the  
 103 generative task. Poyiadzi et al. [21] use density estimators ( $\hat{p} : \mathcal{X} \mapsto [0, 1]$ ) to constrain the  
 104 counterfactual paths. Karimi et al. [11] argue that counterfactuals should comply with the causal  
 105 model that generates the data. All of these different approaches share a common goal: ensuring that  
 106 the generated counterfactuals comply with the true and unobserved DGP. To summarize this broad  
 107 objective, we propose the following definition:

108 **Definition 2.1** (Plausible Counterfactuals). *Let  $\mathcal{X}|\mathbf{y}^*$  denote the true conditional distribution of*  
 109 *samples in the target class  $\mathbf{y}^*$ . Then for  $\mathbf{x}'$  to be considered a plausible counterfactual, we need:*  
 110  $\mathbf{x}' \sim \mathcal{X}|\mathbf{y}^*$ .

111 Note that Definition 2.1 is consistent with the notion of plausible counterfactual paths, since we can  
 112 simply apply it to each counterfactual state along the path.

113 Surrogate models offer an obvious solution to achieve this objective. Unfortunately, surrogates also  
 114 introduce a dependency: the generated explanations no longer depend exclusively on the Black Box  
 115 Model itself, but also on the surrogate model. This is not necessarily problematic if the primary  
 116 objective is not to explain the behaviour of the model but to offer recourse to individuals affected by  
 117 it. It may become problematic even in this context if the dependency turns into a vulnerability. To  
 118 illustrate this point, we have used REVISE [9] with an underfitted VAE to generate the counterfactual  
 119 in the right panel of Figure 2: in this case, the decoder step of the VAE fails to yield plausible values  
 120 ( $\{\mathbf{x}' \leftarrow \mathcal{G}(\mathbf{z})\} \not\sim \mathcal{X}|\mathbf{y}^*$ ) and hence the counterfactual search in the learned latent space is doomed.

## 121 2.2.2 Plausibility through Minimal Predictive Uncertainty

122 Schut et al. [23] show that to meet the plausibility objective we need not explicitly model the input  
123 distribution. Pointing to the undesirable engineering overhead induced by surrogate models, they  
124 propose that we rely on the implicit minimisation of predictive uncertainty instead. Their proposed  
125 methodology solves Equation 1 by greedily applying JSMA in the feature space with standard cross-  
126 entropy loss and no penalty at all. They demonstrate theoretically and empirically that their approach  
127 yields counterfactuals for which the model  $M_\theta$  predicts the target label  $\mathbf{y}^*$  with high confidence.  
128 Provided the model is well-specified, these counterfactuals are plausible. Unfortunately, this idea  
129 hinges on the assumption that the Black Box Model provides well-calibrated predictive uncertainty  
130 estimates.

## 131 2.3 From Fidelity to Model Conformity

132 Above we explained that since Counterfactual Explanations work directly with the Black Box model,  
133 the fidelity of explanations as we defined it earlier is not a concern. This may explain why research  
134 has primarily focused on other desiderata, most notably plausibility (Definition 2.1). Enquiring  
135 about the plausibility of a counterfactual essentially boils down to the following question: ‘Is this  
136 counterfactual consistent with the underlying data’? To introduce this section, we posit a related,  
137 slightly more nuanced question: ‘Is this counterfactual consistent with what the model has learned  
138 about the underlying data’? We will argue that fidelity is not a sufficient evaluation measure to answer  
139 this question and propose a novel way to assess if Counterfactual Explanations conform with model  
140 behaviour.

141 The word *fidelity* stems from the Latin word ‘fidelis’, which means ‘faithful, loyal, trustworthy’  
142 [15]. As we explained in Section 2, model explanations are generally considered faithful if their  
143 corresponding predictions coincide with the predictions made by the model itself. Since this definition  
144 of faithfulness is not useful in the context of Counterfactual Explanations, we propose an adapted  
145 version:

146 **Definition 2.2** (Conformal Counterfactuals). *Let  $\mathcal{X}_\theta|\mathbf{y}^* = p_\theta(x|\mathbf{y}^*)$  denote the conditional distri-*  
147 *bution of  $\mathbf{x}$  in the target class  $\mathbf{y}^*$ , where  $\theta$  denotes the parameters of model  $M_\theta$ . Then for  $\mathbf{x}'$  to be*  
148 *considered a conformal counterfactual, we need:  $\mathbf{x}' \sim \mathcal{X}_\theta|\mathbf{y}^*$ .*

149 In words, conformal counterfactuals conform with what the predictive model has learned about  
150 the input data  $\mathbf{x}$ . Since this definition works with distributional properties, it explicitly accounts  
151 for the multiplicity of explanations we discussed earlier. To assess counterfactuals with respect to  
152 Definition 2.2, we need to be able to quantify the posterior conditional distribution  $p_\theta(\mathbf{x}|\mathbf{y}^*)$ . This is  
153 very much at the core of our proposed methodological framework, which reconciles the notions of  
154 plausibility and model conformity and which we will introduce next.

## 155 3 Methodological Framework

156 The primary objective of this work has been to develop a methodology for generating maximally  
157 plausible counterfactuals under minimal intervention. Our proposed framework is based on the  
158 premise that explanations should be plausible but not plausible at all costs. Energy-Constrained  
159 Conformal Counterfactuals (ECCCo) achieve this goal in two ways: firstly, they rely on the Black  
160 Box itself for the generative task; and, secondly, they involve an approach to predictive uncertainty  
161 quantification that is model-agnostic.

### 162 3.1 Quantifying the Model’s Generative Property

163 Recent work by Grathwohl et al. [7] on Energy Based Models (EBM) has pointed out that there is a  
164 ‘generative model hidden within every standard discriminative model’. The authors show that we can  
165 draw samples from the posterior conditional distribution  $p_\theta(\mathbf{x}|\mathbf{y})$  using Stochastic Gradient Langevin  
166 Dynamics (SGLD). The authors use this insight to train classifiers jointly for the discriminative task  
167 using standard cross-entropy and the generative task using SGLD. They demonstrate empirically that  
168 among other things this improves predictive uncertainty quantification for discriminative models.  
169 Our findings in this work suggest that Joint Energy Models (JEM) also tend to yield more plausible

170 Counterfactual Explanations. Based on the definition of plausible counterfactuals (Definition 2.1)  
 171 this is not surprising.

172 Crucially for our purpose, one can apply their proposed sampling strategy during inference to  
 173 essentially any standard discriminative model. Even models that are not explicitly trained for the joint  
 174 objective learn about the distribution of inputs  $X$  by learning to make conditional predictions about  
 175 the output  $y$ . We can leverage this observation to quantify the generative property of the Black Box  
 176 model itself. In particular, note that if we fix  $\mathbf{y}$  to our target value  $\mathbf{y}^*$ , we can sample from  $p_\theta(\mathbf{x}|\mathbf{y}^*)$   
 177 using SGLD as follows,

$$\mathbf{x}_{j+1} \leftarrow \mathbf{x}_j - \frac{\epsilon^2}{2} \mathcal{E}(\mathbf{x}_j|\mathbf{y}^*) + \epsilon \mathbf{r}_j, \quad j = 1, \dots, J \quad (2)$$

178 where  $\mathbf{r}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  is the stochastic term and the step-size  $\epsilon$  is typically polynomially decayed.  
 179 The term  $\mathcal{E}(\mathbf{x}_j|\mathbf{y}^*)$  denotes the energy function where we use  $\mathcal{E}(\mathbf{x}_j|\mathbf{y}^*) = -M_\theta(\mathbf{x}_j)[\mathbf{y}^*]$ , that is the  
 180 negative logit corresponding to the target class label  $\mathbf{y}^*$ . Generating multiple samples in this manner  
 181 yields an empirical distribution  $\hat{\mathcal{X}}_\theta|\mathbf{y}^*$  that we use in our search for plausible counterfactuals, as  
 182 discussed in more detail below. Appendix A provides additional implementation details for any tasks  
 183 related to energy-based modelling.

### 184 3.2 Quantifying the Model’s Predictive Uncertainty

185 To quantify the model’s predictive uncertainty we use Conformal Prediction (CP), an approach that  
 186 has recently gained popularity in the Machine Learning community [3, 14]. Crucially for our intended  
 187 application, CP is model-agnostic and can be applied during inference without placing any restrictions  
 188 on model training. Intuitively, CP works under the premise of turning heuristic notions of uncertainty  
 189 into rigorous uncertainty estimates by repeatedly sifting through the training data or a dedicated  
 190 calibration dataset. Conformal classifiers produce prediction sets for individual inputs that include all  
 191 output labels that can be reasonably attributed to the input. These sets tend to be larger for inputs that  
 192 do not conform with the training data and are therefore characterized by high predictive uncertainty.

193 In order to generate counterfactuals that are associated with low predictive uncertainty, we use a  
 194 smooth set size penalty introduced by Stutz et al. [25] in the context of conformal training:

$$\Omega(C_\theta(\mathbf{x}; \alpha)) = \max \left( 0, \sum_{\mathbf{y} \in \mathcal{Y}} C_{\theta, \mathbf{y}}(\mathbf{x}_i; \alpha) - \kappa \right) \quad (3)$$

195 Here,  $\kappa \in \{0, 1\}$  is a hyper-parameter and  $C_{\theta, \mathbf{y}}(\mathbf{x}_i; \alpha)$  can be interpreted as the probability of label  
 196  $\mathbf{y}$  being included in the prediction set.

197 In order to compute this penalty for any Black Box Model we merely need to perform a single  
 198 calibration pass through a holdout set  $\mathcal{D}_{\text{cal}}$ . Arguably, data is typically abundant and in most  
 199 applications, practitioners tend to hold out a test data set anyway. Consequently, CP removes the  
 200 restriction on the family of predictive models, at the small cost of reserving a subset of the available  
 201 data for calibration. This particular case of conformal prediction is referred to as Split Conformal  
 202 Prediction (SCP) as it involves splitting the training data into a proper training dataset and a calibration  
 203 dataset. Details concerning our implementation of Conformal Prediction can be found in Appendix B.

### 204 3.3 Energy-Constrained Conformal Counterfactuals (ECCCo)

205 Our framework for generating ECCCos combines the ideas introduced in the previous two subsections.  
 206 Formally, we extend Equation 1 as follows,

$$\begin{aligned} \mathbf{Z}' = \arg \min_{\mathbf{Z}' \in \mathcal{Z}^M} \{ & \text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^*) + \lambda_1 \text{dist}(f(\mathbf{Z}'), \mathbf{x}) \\ & + \lambda_2 \text{dist}(f(\mathbf{Z}'), \hat{\mathbf{x}}_\theta) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}'); \alpha)) \} \end{aligned} \quad (4)$$

207 where  $\hat{\mathbf{x}}_\theta$  denotes samples generated using SGLD (Equation 2) and  $\text{dist}(\cdot)$  is a generic term for a  
 208 distance metric. Our default choice for  $\text{dist}(\cdot)$  is the Manhattan Distance since it enforces sparsity.

209 The first two terms in Equation 4 correspond to the counterfactual search objective defined in Wachter  
 210 et al. [29] which merely penalises the distance of counterfactuals from their factual values. The  
 211 additional two penalties in ECCCo ensure that counterfactuals conform with the model’s generative  
 212 property and lead to minimally uncertain predictions, respectively. The hyperparameters  $\lambda_1, \dots, \lambda_3$   
 213 can be used to balance the different objectives: for example, we may choose to incur larger deviations  
 214 from the factual in favour of conformity with the model’s generative property by choosing lower  
 215 values of  $\lambda_1$  and relatively higher values of  $\lambda_2$ . Figure 4 illustrates this balancing act for an example  
 216 involving synthetic data: vector fields indicate the direction of gradients with respect to the different  
 217 components our proposed objective function (Equation 4).

218 The entire procedure for Generating ECCCos is described in Algorithm 1. For the sake of simplicity  
 219 and without loss of generality, we limit our attention to generating a single counterfactual  $\mathbf{x}' = f(\mathbf{z}')$   
 220 where in contrast to Equation 4  $\mathbf{z}'$  denotes a 1-dimensional array containing a single counterfactual  
 221 state. That state is initialized by passing the factual  $\mathbf{x}$  through the encoder  $f^{-1}$  which in our case  
 222 corresponds to a simple feature transformer, rather than the encoder part of VAE as in REVISE [9].

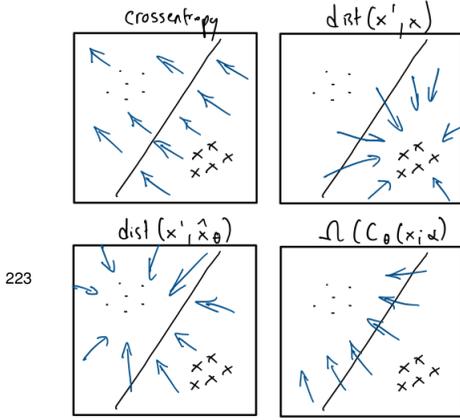


Figure 3: Vector fields indicating the direction of gradients with respect to the different components of the ECCCo objective (Equation 4).

Algorithm 1: Generating ECCCos (For more details, see Appendix C)

**Input:**  $\mathbf{x}, \mathbf{y}^*, M_\theta, f, \Lambda, \alpha, \mathcal{D}, T, \eta, m, M$   
 where  $M_\theta(\mathbf{x}) \neq \mathbf{y}^*$

**Output:**  $\mathbf{x}'$

- 1: Initialize  $\mathbf{z}' \leftarrow f^{-1}(\mathbf{x})$
- 2: Generate buffer  $\mathcal{B}$  of  $M$  conditional samples  $\hat{\mathbf{x}}_\theta | \mathbf{y}^*$  using SGLD (Equation 2)
- 3: Run SCP for  $M_\theta$  using  $\mathcal{D}$
- 4: Initialize  $t \leftarrow 0$
- 5: **while** not converged or  $t < T$  **do**
- 6:    $\hat{\mathbf{x}}_{\theta,t} \leftarrow \text{rand}(\mathcal{B}, m)$
- 7:    $\mathbf{z}' \leftarrow \mathbf{z}' - \eta \nabla_{\mathbf{z}'} \mathcal{L}(\mathbf{z}', \mathbf{y}^*, \hat{\mathbf{x}}_{\theta,t})$
- 8:    $t \leftarrow t + 1$
- 9: **end while**
- 10:  $\mathbf{x}' \leftarrow f(\mathbf{z}')$

## 224 4 Evaluation Framework

### 225 4.1 Evaluation Measures

226 Above we have defined plausibility (2.1) and conformity (2.2) for Counterfactual Explanations.  
 227 In this subsection, we introduce evaluation measures that facilitate a quantitative evaluation of  
 228 counterfactuals for these objectives.

229 Firstly, in order to assess the plausibility of counterfactuals we adapt the implausibility metric  
 230 proposed in Guidotti [8]. The authors propose to evaluate plausibility in terms of the distance of the  
 231 counterfactual  $\mathbf{x}'$  from its nearest neighbour in the target class  $\mathbf{y}^*$ : the smaller this distance, the more  
 232 plausible the counterfactual. Instead of focusing only on the nearest neighbour of  $\mathbf{x}'$ , we suggest  
 233 computing the average over distances from multiple (possibly all) observed instances in the target  
 234 class. Formally, for a single counterfactual, we have:

$$\text{impl} = \frac{1}{|\mathbf{x} \in \mathcal{X} | \mathbf{y}^*|} \sum_{\mathbf{x} \in \mathcal{X} | \mathbf{y}^*} \text{dist}(\mathbf{x}', \mathbf{x}) \quad (5)$$

235 This measure is straightforward to compute and should be less sensitive to outliers in the target class  
 236 than the one based on the nearest neighbour. It also gives rise to a very similar evaluation measure for  
 237 conformity. We merely swap out the subsample of individuals in the target class for the empirical  
 238 distribution of generated conditional samples:

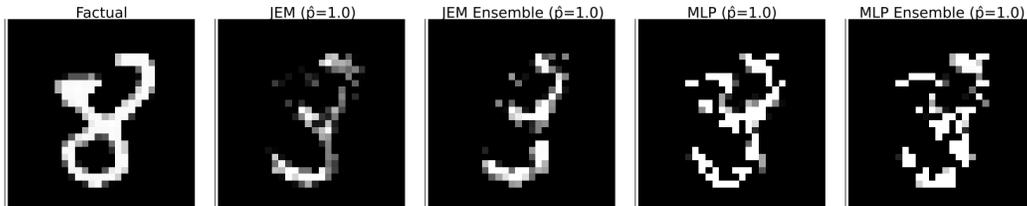


Figure 4: ECCCos from Black Boxes. Counterfactuals for turning an 8 (eight) into a 3 (three): original image (left);

$$\text{conf} = \frac{1}{|\mathbf{x} \in \mathcal{X}_\theta | \mathbf{y}^*|} \sum_{\mathbf{x} \in \mathcal{X}_\theta | \mathbf{y}^*} \text{dist}(\mathbf{x}', \mathbf{x}) \quad (6)$$

239 As noted by Guidotti [8], these distance-based measures are simplistic and more complex alternative  
 240 measures may ultimately be more appropriate for the task. For example, we considered using statisti-  
 241 cal divergence measures instead. This would involve generating not one but many counterfactuals and  
 242 comparing the generated empirical distribution to the target distributions in Definitions 2.1 and 2.2.  
 243 While this approach is potentially more rigorous, generating enough counterfactuals is not always  
 244 practical.

## 245 5 Experiments

- 246 • BatchNorm does not seem compatible with JEM
- 247 • Coverage and temperature impacts CCE in somewhat unpredictable ways
- 248 • It seems that models that are not explicitly trained for generative task, still learn it implicitly
- 249 • Batch size seems to impact quality of generated samples (at inference, but not so much  
 250 during JEM training)
- 251 • ECCCo is sensitive to optimizer (Adam works well), learning rate and distance metric (l1  
 252 currently only one that works)
- 253 • SGLD takes time
- 254 • REVISE has benefit of lower dimensional space
- 255 • For MNIST it seems that ECCCo is better at reducing pixel values than increasing them  
 256 (better at erasing than writing)

## 257 6 Discussion

258 Consistent with the findings in Schut et al. [23], we have demonstrated that predictive uncertainty  
 259 estimates can be leveraged to generate plausible counterfactuals. Interestingly, Schut et al. [23]  
 260 point out that this finding — as intuitive as it is — may be linked to a positive connection between  
 261 the generative task and predictive uncertainty quantification. In particular, Grathwohl et al. [7]  
 262 demonstrate that their proposed method for integrating the generative objective in training yields  
 263 models that have improved predictive uncertainty quantification. Since neither Schut et al. [23] nor  
 264 we have employed any surrogate generative models, our findings seem to indicate that the positive  
 265 connection found in Grathwohl et al. [7] is bidirectional.

## 266 References

- 267 [1] Patrick Altmeyer. Conformal Prediction in Julia. URL [https://www.paltmeyer.com/blog/  
 268 posts/conformal-prediction/](https://www.paltmeyer.com/blog/posts/conformal-prediction/).
- 269 [2] Patrick Altmeyer, Giovan Angela, Aleksander Buszydlík, Karol Dobiczek, Arie van Deursen,  
 270 and Cynthia Liem. Endogenous Macrodynamics in Algorithmic Recourse. In *First IEEE  
 271 Conference on Secure and Trustworthy Machine Learning, 2023*.

- 272 [3] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction  
273 and distribution-free uncertainty quantification. 2021.
- 274 [4] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling,  
275 and Barbara Hammer. Evaluating Robustness of Counterfactual Explanations. Technical report,  
276 arXiv. URL <http://arxiv.org/abs/2103.02354>. arXiv:2103.02354 [cs] type: article.
- 277 [5] Ann-Kathrin Dombrowski, Jan E Gerken, and Pan Kessel. Diffeomorphic explanations with  
278 normalizing flows. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and*  
279 *Explicit Likelihood Models*, 2021.
- 280 [6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial  
281 examples. 2014.
- 282 [7] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad  
283 Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should  
284 treat it like one. March 2020. URL <https://openreview.net/forum?id=HkxzxONtDB>.
- 285 [8] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and  
286 benchmarking. ISSN 1573-756X. doi: 10.1007/s10618-022-00831-6. URL <https://doi.org/10.1007/s10618-022-00831-6>.
- 288 [9] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh.  
289 Towards realistic individual recourse and actionable explanations in black-box decision making  
290 systems. 2019.
- 291 [10] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of  
292 algorithmic recourse: Definitions, formulations, solutions, and prospects. 2020.
- 293 [11] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From  
294 counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on*  
295 *Fairness, Accountability, and Transparency*, pages 353–362, 2021.
- 296 [12] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In  
297 *Proceedings of the 31st International Conference on Neural Information Processing Systems*,  
298 pages 4768–4777, 2017.
- 299 [13] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving Causal Constraints in Counter-  
300 factual Explanations for Machine Learning Classifiers. Technical report, arXiv. URL  
301 <http://arxiv.org/abs/1912.03277>. arXiv:1912.03277 [cs, stat] type: article.
- 302 [14] Valery Manokhin. Awesome conformal prediction.
- 303 [15] Merriam-Webster. "fidelity". URL [https://www.merriam-webster.com/dictionary/](https://www.merriam-webster.com/dictionary/fidelity)  
304 [fidelity](https://www.merriam-webster.com/dictionary/fidelity).
- 305 [16] Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020.
- 306 [17] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning  
307 classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference*  
308 *on Fairness, Accountability, and Transparency*, pages 607–617, 2020.
- 309 [18] Kevin P. Murphy. *Probabilistic machine learning: Advanced topics*. MIT Press.
- 310 [19] Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji  
311 Kasneci. Carla: A python library to benchmark algorithmic recourse and counterfactual  
312 explanation algorithms. 2021.
- 313 [20] Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu  
314 Lakkaraju. Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and  
315 Robustness in Algorithmic Recourse. *arXiv preprint arXiv:2203.06768*, 2022.
- 316 [21] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijn De Bie, and Peter Flach. FACE:  
317 Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Confer-*  
318 *ence on AI, Ethics, and Society*, pages 344–350, 2020.

- 319 [22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining  
320 the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International*  
321 *Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.
- 322 [23] Lisa Schut, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al.  
323 Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic  
324 and Aleatoric Uncertainties. In *International Conference on Artificial Intelligence and Statistics*,  
325 pages 1756–1764. PMLR, 2021.
- 326 [24] Thomas Spooner, Danial Dervovic, Jason Long, Jon Shepard, Jiahao Chen, and Daniele Maga-  
327 zzeni. Counterfactual Explanations for Arbitrary Regression Models. 2021.
- 328 [25] David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning  
329 Optimal Conformal Classifiers. May 2022. URL <https://openreview.net/forum?id=t80-4LKfVx>.
- 331 [26] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards Robust and Reliable  
332 Algorithmic Recourse. 2021.
- 333 [27] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In  
334 *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19,  
335 2019.
- 336 [28] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine  
337 learning: A review. 2020.
- 338 [29] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without  
339 opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.
- 340 [30] Andrew Gordon Wilson. The case for Bayesian deep learning. 2020.

## 341 Appendices

### 342 A JEM

343 While  $\mathbf{x}_J$  is only guaranteed to distribute as  $p_\theta(\mathbf{x}|\mathbf{y}^*)$  if  $\epsilon \rightarrow 0$  and  $J \rightarrow \infty$ , the bias introduced for  
344 a small finite  $\epsilon$  is negligible in practice [18, 7]. While Grathwohl et al. [7] use Equation 2 during  
345 training, we are interested in applying the conditional sampling procedure in a post hoc fashion to  
346 any standard discriminative model.

### 347 B Conformal Prediction

348 The fact that conformal classifiers produce set-valued predictions introduces a challenge: it is not  
349 immediately obvious how to use such classifiers in the context of gradient-based counterfactual  
350 search. Put differently, it is not clear how to use prediction sets in Equation 1. Fortunately, Stutz et al.  
351 [25] have recently proposed a framework for Conformal Training that also hinges on differentiability.  
352 Specifically, they show how Stochastic Gradient Descent can be used to train classifiers not only  
353 for the discriminative task but also for additional objectives related to Conformal Prediction. One  
354 such objective is *efficiency*: for a given target error rate *alpha*, the efficiency of a conformal classifier  
355 improves as its average prediction set size decreases. To this end, the authors introduce a smooth set  
356 size penalty defined in Equation 3

357 Formally, it is defined as  $C_{\theta, \mathbf{y}}(\mathbf{x}_i; \alpha) := \sigma((s(\mathbf{x}_i, \mathbf{y}) - \alpha)T^{-1})$  for  $\mathbf{y} \in \mathcal{Y}$  where  $\sigma$  is the sigmoid  
358 function and  $T$  is a hyper-parameter used for temperature scaling [25].

359 Intuitively, CP works under the premise of turning heuristic notions of uncertainty into rigorous  
360 uncertainty estimates by repeatedly sifting through the data. It can be used to generate prediction  
361 intervals for regression models and prediction sets for classification models [1]. Since the literature  
362 on CE and AR is typically concerned with classification problems, we focus on the latter. A particular  
363 variant of CP called Split Conformal Prediction (SCP) is well-suited for our purposes because it  
364 imposes only minimal restrictions on model training.

365 Specifically, SCP involves splitting the data  $\mathcal{D}_n = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1, \dots, n}$  into a proper training set  $\mathcal{D}_{\text{train}}$   
366 and a calibration set  $\mathcal{D}_{\text{cal}}$ . The former is used to train the classifier in any conventional fashion.  
367 The latter is then used to compute so-called nonconformity scores:  $\mathcal{S} = \{s(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in \mathcal{D}_{\text{cal}}}$  where  
368  $s : (\mathcal{X}, \mathcal{Y}) \mapsto \mathbb{R}$  is referred to as *score function*. In the context of classification, a common choice for  
369 the score function is just  $s_i = 1 - M_\theta(\mathbf{x}_i)[\mathbf{y}_i]$ , that is one minus the softmax output corresponding  
370 to the observed label  $\mathbf{y}_i$  [3].

371 Finally, classification sets are formed as follows,

$$C_\theta(\mathbf{x}_i; \alpha) = \{\mathbf{y} : s(\mathbf{x}_i, \mathbf{y}) \leq \hat{q}\} \quad (7)$$

372 where  $\hat{q}$  denotes the  $(1 - \alpha)$ -quantile of  $\mathcal{S}$  and  $\alpha$  is a predetermined error rate. As the size of the  
373 calibration set increases, the probability that the classification set  $C(\mathbf{x}_{\text{test}})$  for a newly arrived sample  
374  $\mathbf{x}_{\text{test}}$  does not cover the true test label  $\mathbf{y}_{\text{test}}$  approaches  $\alpha$  [3].

375 Observe from Equation 7 that Conformal Prediction works on an instance-level basis, much like  
376 Counterfactual Explanations are local. The prediction set for an individual instance  $\mathbf{x}_i$  depends only  
377 on the characteristics of that sample and the specified error rate. Intuitively, the set is more likely  
378 to include multiple labels for samples that are difficult to classify, so the set size is indicative of  
379 predictive uncertainty. To see why this effect is exacerbated by small choices for  $\alpha$  consider the case  
380 of  $\alpha = 0$ , which requires that the true label is covered by the prediction set with probability equal to  
381 one.

## 382 C Conformal Prediction

### 383 A Submission of papers to NeurIPS 2023

384 Please read the instructions below carefully and follow them faithfully.

#### 385 A Style

386 Papers to be submitted to NeurIPS 2023 must be prepared according to the instructions presented  
387 here. Papers may only be up to **nine** pages long, including figures. Additional pages *containing only*  
388 *acknowledgments and references* are allowed. Papers that exceed the page limit will not be reviewed,  
389 or in any other way considered for presentation at the conference.

390 The margins in 2023 are the same as those in previous years.

391 Authors are required to use the NeurIPS L<sup>A</sup>T<sub>E</sub>X style files obtainable at the NeurIPS website as  
392 indicated below. Please make sure you use the current files and not previous versions. Tweaking the  
393 style files may be grounds for rejection.

#### 394 B Retrieval of style files

395 The style files for NeurIPS and other conference information are available on the website at

396 <http://www.neurips.cc/>

397 The file `neurips_2023.pdf` contains these instructions and illustrates the various formatting re-  
398 quirements your NeurIPS paper must satisfy.

399 The only supported style file for NeurIPS 2023 is `neurips_2023.sty`, rewritten for L<sup>A</sup>T<sub>E</sub>X 2<sub>ε</sub>.  
400 **Previous style files for L<sup>A</sup>T<sub>E</sub>X 2.09, Microsoft Word, and RTF are no longer supported!**

401 The L<sup>A</sup>T<sub>E</sub>X style file contains three optional arguments: `final`, which creates a camera-ready copy,  
402 `preprint`, which creates a preprint for submission to, e.g., arXiv, and `nonatbib`, which will not  
403 load the `natbib` package for you in case of package clash.

404 **Preprint option** If you wish to post a preprint of your work online, e.g., on arXiv, using the  
405 NeurIPS style, please use the `preprint` option. This will create a nonanonymized version of your  
406 work with the text “Preprint. Work in progress.” in the footer. This version may be distributed as you

407 see fit, as long as you do not say which conference it was submitted to. Please **do not** use the `final`  
408 option, which should **only** be used for papers accepted to NeurIPS.

409 At submission time, please omit the `final` and `preprint` options. This will anonymize your  
410 submission and add line numbers to aid review. Please do *not* refer to these line numbers in your  
411 paper as they will be removed during generation of camera-ready copies.

412 The file `neurips_2023.tex` may be used as a “shell” for writing your paper. All you have to do is  
413 replace the author, title, abstract, and text of the paper with your own.

414 The formatting instructions contained in these style files are summarized in Sections B, C, and D  
415 below.

## 416 **B General formatting instructions**

417 The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long.  
418 The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points.  
419 Times New Roman is the preferred typeface throughout, and will be selected for you by default.  
420 Paragraphs are separated by  $\frac{1}{2}$  line space (5.5 points), with no indentation.

421 The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal  
422 rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow  $\frac{1}{4}$  inch  
423 space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the  
424 page.

425 For the final version, authors’ names are set in boldface, and each name is centered above the  
426 corresponding address. The lead author’s name is to be listed first (left-most), and the co-authors’  
427 names (if different address) are set to follow. If there is only one co-author, list both author and  
428 co-author side by side.

429 Please pay special attention to the instructions in Section D regarding figures, tables, acknowledg-  
430 ments, and references.

## 431 **C Headings: first level**

432 All headings should be lower case (except for first word and proper nouns), flush left, and bold.

433 First-level headings should be in 12-point type.

### 434 **A Headings: second level**

435 Second-level headings should be in 10-point type.

#### 436 **A.1 Headings: third level**

437 Third-level headings should be in 10-point type.

438 **Paragraphs** There is also a `\paragraph` command available, which sets the heading in bold, flush  
439 left, and inline with the text, with the heading followed by 1 em of space.

## 440 **D Citations, figures, tables, references**

441 These instructions apply to everyone.

### 442 **A Citations within the text**

443 The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as  
444 long as you maintain internal consistency. As to the format of the references themselves, any style is  
445 acceptable as long as it is used consistently.

446 The documentation for `natbib` may be found at

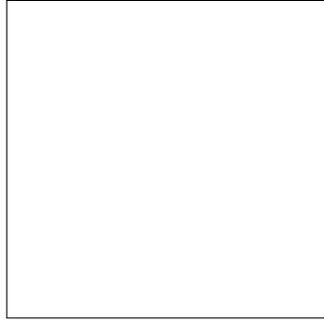


Figure 5: Sample figure caption.

447 `http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

448 Of note is the command `\citet`, which produces citations appropriate for use in inline text. For  
449 example,

450 `\citet{hasselmo}` investigated\dots

451 produces

452 Hasselmo, et al. (1995) investigated...

453 If you wish to load the `natbib` package with options, you may add the following before loading the  
454 `neurips_2023` package:

455 `\PassOptionsToPackage{options}{natbib}`

456 If `natbib` clashes with another package you load, you can add the optional argument `nonatbib`  
457 when loading the style file:

458 `\usepackage[nonatbib]{neurips_2023}`

459 As submission is double blind, refer to your own published work in the third person. That is, use “In  
460 the previous work of Jones et al. [4],” not “In our previous work [4].” If you cite your other papers  
461 that are not widely available (e.g., a journal paper under review), use anonymous author names in the  
462 citation, e.g., an author of the form “A. Anonymous” and include a copy of the anonymized paper in  
463 the supplementary material.

## 464 **B Footnotes**

465 Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number<sup>1</sup>  
466 in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote  
467 with a horizontal rule of 2 inches (12 picas).

468 Note that footnotes are properly typeset *after* punctuation marks.<sup>2</sup>

## 469 **C Figures**

470 All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction.  
471 The figure number and caption always appear after the figure. Place one line space before the figure  
472 caption and one line space after the figure. The figure caption should be lower case (except for first  
473 word and proper nouns); figures are numbered consecutively.

474 You may use color figures. However, it is best for the figure captions and the paper body to be legible  
475 if the paper is printed in either black/white or in color.

---

<sup>1</sup>Sample of the first footnote.

<sup>2</sup>As in this example.

Table 1: Sample table title

Part		
Name	Description	Size ( $\mu\text{m}$ )
Dendrite	Input terminal	$\sim 100$
Axon	Output terminal	$\sim 10$
Soma	Cell body	up to $10^6$

476 **D Tables**

477 All tables must be centered, neat, clean and legible. The table number and title always appear before  
478 the table. See Table 1.

479 Place one line space before the table title, one line space after the table title, and one line space after  
480 the table. The table title must be lower case (except for first word and proper nouns); tables are  
481 numbered consecutively.

482 Note that publication-quality tables *do not contain vertical rules*. We strongly suggest the use of the  
483 `booktabs` package, which allows for typesetting high-quality, professional tables:

484 <https://www.ctan.org/pkg/booktabs>

485 This package was used to typeset Table 1.

486 **E Math**

487 Note that display math in bare TeX commands will not create correct line numbers for sub-  
488 mission. Please use LaTeX (or AMSTeX) commands for unnumbered display math. (You  
489 really shouldn't be using  $\$$  anyway; see <https://tex.stackexchange.com/questions/503/why-is-preferable-to> and <https://tex.stackexchange.com/questions/40492/what-are-the-differences-between-align-equation-and-displaymath> for more infor-  
491 mation.)

493 **F Final instructions**

494 Do not change any aspects of the formatting parameters in the style files. In particular, do not modify  
495 the width or length of the rectangle the text should fit into, and do not change font sizes (except  
496 perhaps in the **References** section; see below). Please note that pages should be numbered.

497 **E Preparing PDF files**

498 Please prepare submission files with paper size "US Letter," and not, for example, "A4."

499 Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or  
500 Embedded TrueType fonts. Here are a few instructions to achieve this.

- 501 • You should directly generate PDF files using `pdflatex`.
- 502 • You can check which fonts a PDF files uses. In Acrobat Reader, select the menu  
503 Files>Document Properties>Fonts and select Show All Fonts. You can also use the program  
504 `pdf-fonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- 505 • `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- 506 • The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS  
507 Fonts:

508 `\usepackage{amsfonts}`

509 followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for  $\mathbb{R}$ ,  $\mathbb{N}$  or  $\mathbb{C}$ . You can also  
510 use the following workaround for reals, natural and complex:

```
511 \newcommand{\RR}{\mathbb{R}} %real numbers
512 \newcommand{\Nat}{\mathbb{N}} %natural numbers
513 \newcommand{\CC}{\mathbb{C}} %complex numbers
```

514 Note that `amsfonts` is automatically loaded by the `amssymb` package.

515 If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

## 516 A Margins in L<sup>A</sup>T<sub>E</sub>X

517 Most of the margin problems come from figures positioned by hand using `\special` or other  
518 commands. We suggest using the command `\includegraphics` from the `graphicx` package.  
519 Always specify the figure width as a multiple of the line width as in the example below:

```
520 \usepackage[pdftex]{graphicx} ...
521 \includegraphics[width=0.8\linewidth]{myfile.pdf}
```

522 See Section 4.4 in the graphics bundle documentation ([http://mirrors.ctan.org/macros/](http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf)  
523 [latex/required/graphics/grfguide.pdf](http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf))

524 A number of width problems arise when L<sup>A</sup>T<sub>E</sub>X cannot properly hyphenate a line. Please give LaTeX  
525 hyphenation hints using the `\-` command when necessary.

## 526 F Supplementary Material

527 Authors may wish to optionally include extra information (complete proofs, additional experiments  
528 and plots) in the appendix. All such materials should be part of the supplemental material (submitted  
529 separately) and should NOT be included in the main submission.

## 530 References

531 References follow the acknowledgments in the camera-ready paper. Use unnumbered first-level  
532 heading for the references. Any choice of citation style is acceptable as long as you are consistent. It  
533 is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the  
534 Reference section does not count towards the page limit.

535 [1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In  
536 G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp.  
537 609–616. Cambridge, MA: MIT Press.

538 [2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the*  
539 *General NEural Simulation System*. New York: TELOS/Springer-Verlag.

540 [3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent  
541 synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.