Games for Survival Analysis

Kaplan Meier *
Department of Computer Science
Cranberry-Lemon University
Pittsburgh, PA 15213
hippo@cs.cranberry-lemon.edu

Abstract

Deep discrete models have become a common choice for survival analysis. Though training is usually done through maximum likelihood, practitioners evaluate under other criteria, such as binary classification losses at a set of chosen time horizons (e.g. Brier score (BS) and Bernoulli log likelihood (BLL) for 5 year prediction) Due to finite data, these models may have poor BS and BLL. However, maximum likelihood does not directly optimize these criteria. Doing so requires re-weighting by the censorship distribution which is generally unknown. To resolve this dilemma, we introduce *Inverse-Weighted Survival Games*. In these non-adversarial games, failure and censoring models simultaneously optimize a set of losses. Each loss is built from re-weighted estimates where the models account for censoring in each other's losses. When one model is correct, the other directly optimizes the criteria that would be computed under no censoring. When the loss is proper, we show that the games always have the true failure and censoring distributions as a stationary point. This means models in the game do not leave the correct distributions once reached. When the game is based on BS, the true distributions are the *only* stationary point. We show that these games optimize the true uncensored BS on simulations and then apply these principles on real world data.

1 Introduction

Survival analysis is the modeling of time-to-event distributions and is widely used in healthcare to predict time from diagnosis to death, risk of disease recurrence, and changes in level of care. In survival data, events, known as *failures* may be *censored*, i.e. only a time range is known. Under right-censoring, we only observe a lower-bound on some failure times, for instance when a patient leaves a study before failing [Huang et al., 2004]. Though the usual goal is to model failures, discarding censored points biases estimates. In this work, we revisit the role of censoring during training.

Under independent censoring, likelihood training only requires evaluating a failure model on observed failure and censoring times [Kalbfleisch and Prentice, 2002]. Most practitioners make these assumptions and avoid the nuisance of modeling censoring explicitly. However, we show that, even under these assumptions, modeling censoring can help with modeling and evaluating failure distributions.

Deep models have dominated recent survival work [Ranganath et al., 2016, Alaa and van der Schaar, 2017, Katzman et al., 2018, Kvamme et al., 2019, Zhong and Tibshirani, 2019, Goldstein et al., 2020]. We focus on deep *discrete-time* models, which have become common even when data are continuous [Yu et al., 2011, Lee et al., 2018, Fotso, 2018, Lee et al., 2019, Ren et al., 2019, Kvamme and Borgan, 2019b, Kamran and Wiens, 2021, Sloma et al., 2021] because they can borrow classification architectures and approximate continuous densities well with enough bins [Miscouridou et al., 2018].

*todo further information about author (webpage, alternative address)—*not* funding.

Preprint. Under review.

Though training is often based on maximum likelihood, survival analysis is full of other evaluation criteria [Haider et al., 2020]. The Brier score (BS) and Bernoulli log likelihood (BLL) are classification losses adapted for survival by treating the model as a binary classifier at various time horizons (*will the event occur before or after 5 years?*) [Kvamme and Borgan, 2019b, Lee et al., 2019, Steingrimsson and Morrison, 2020]. These metrics can also be motivated by calibration (Section 3) which is valuable because survival probabilities are used to communicate risk [Sullivan et al., 2004].

The story of deep models and survival evaluations is not simple. Due to finite data, models may have poor BS and BLL. However maximum likelihood does not directly optimize these criteria. Due to missingness, doing so requires re-weighting using inverse probability of censor-weighting (IPCW) [Van der Laan et al., 2003]. However, IPCW requires the censorship distribution, which is generally unknown. But since censoring events are themselves censored by observed failures, learning the censoring distribution is also a challenging task. This poses a re-weighting dilemma: each model is required for training the other model under these criteria, but neither are known.

To resolve this dilemma, we introduce *Inverse-Weighted Survival Games*. In these non-adversarial games, a failure model and censoring model respectively optimize a set of losses. These losses are built from IPCW estimates where the two models re-weight each other's loss. Throughout training, the models learn to properly re-weighting to account for censoring in each other's tasks: when one distribution is correct, the other directly optimizes consistent estimates of its uncensored loss.

When the loss is *proper* (e.g. BS, BLL) [Gneiting and Raftery, 2007], we show that the games have the true failure and censoring distributions as a stationary point. This means the models in the game do not leave the correct distributions once reached. We then show that when the game is based on BS, these distributions are the *only* stationary point. Using the games, we optimize the uncensored BS on simulations and apply these principles on real world cancer and ill-patient data.

2 Notation and background on IPCW

Notation. Let T be a failure time with CDF F, survival function $\overline{F} = 1 - F$, and model F_{θ_T}. Let C be a censoring time with CDF G, $\overline{G} = 1 - G$, and model G_{θ_C}. Let X denote features. Under right-censoring, $U = \min(T, C)$, $\Delta = \mathbb{1}[T \leq C]$ and we observe (X, U, Δ). Let $\overline{G}(t^-)$ denote $P(C \geq t)$. For discrete models over K times, let $\theta_{Tt} = P_\theta(T = t)$ and $\theta_T = \{\theta_{Tk}\}_{k=1}^{K-1}$, and likewise for C and θ_C.

Assumptions. We assume i.i.d. data and random censoring: $T \perp\!\!\!\perp C \mid X$ [Kalbfleisch and Prentice, 2002]. We also require the censoring positivity assumption [Gerds et al., 2013]. Let $f = dF$. Then:

$$\forall x \; \exists \epsilon_x \qquad \text{s.t.} \qquad \forall t \in \{t \mid f(t|x) > 0\}, \quad \overline{G}(t^-|x) \geq \epsilon_x > 0. \tag{1}$$

This means we observe all events with positive probability. Though "censoring" typically refers to missingness of failure times, in this work we model censoring too, a survival modeling task where censoring times are censored by observed failures. In order to observe censoring events properly, we also require positivity to hold with the roles of F and G reversed (Appendix B).

Models. We focus on deep discrete models like those studied in Lee et al. [2018], Kvamme and Borgan [2019b]. The model maps inputs X to a categorical distribution over times. When the observations continuous, a discretization scheme is necessary. Following Kvamme and Borgan [2019b], Goldstein et al. [2020], we set bins to correspond to quantiles of observed times. We represent all times in an interval by its lower boundary.

IPCW Estimators. Inverse probability of censor-weighting (IPCW) is a method for estimation under censoring [Van der Laan et al., 2003, Bang and Robins, 2005]. Consider the marginal mean $\mathbb{E}[T]$. IPCW reformulates such expectations in terms of observed data. Using IPCW, we can show that:

$$\mathbb{E}[T] = \mathbb{E}\left[\frac{\mathbb{1}[T \leq C]}{\overline{G}(T^-|X)} T\right] = \mathbb{E}\left[\frac{\Delta U}{\overline{G}(U^-|X)}\right]$$

We derive this in Appendix C. The second equality holds because $\Delta = 1 \implies U = T$ and means we can identify $\mathbb{E}[T]$ from observed data, provided that we know G and that random censoring and positivity (Equation (1)) hold.

3 Time-dependent survival evaluations

Brier score (BS) [Brier and Allen, 1951] is a *strictly proper* scoring rule for classification tasks, i.e. it takes its minimum value only for the true data-generating distribution [Gneiting and Raftery, 2007]. The BS is often adapted for survival modeling [Lee et al., 2019, Kvamme et al., 2019, Haider et al., 2020]. For time t, it computes squared error between the CDF and true event status at t, turning survival analysis into a classification problem at a given time horizon:

$$\text{BS}(t;\theta) = \mathbb{E}\left[\left(F_{\theta_T}(t \mid X) - \mathbb{1}\left[T \leq t\right]\right)^2\right] \tag{2}$$

BS is often used as a proxy for marginal calibration error [Kumar et al., 2018, Lee et al., 2019], which measures differences between cumulative distribution function (CDF) levels $\alpha \in [0, 1]$ and observed proportions of datapoints with $F_\theta(T|X) < \alpha$ [Demler et al., 2015]. BS can be decomposed into such a calibration term plus a discriminative mean squared error term [DeGroot and Fienberg, 1983].

Unfortunately one cannot compute BS unmodified since $\mathbb{1}\left[T \leq t\right]$ is unobserved for a point censored before t. IPCW BS [Graf et al., 1999, Gerds and Schumacher, 2006] estimates $\text{BS}(t)$ under censoring:

$$\text{BS}(t;\theta) = \mathbb{E}\left[\frac{\overline{F}_{\theta_T}(t \mid X)^2 \Delta \mathbb{1}\left[U \leq t\right]}{\overline{G}(U^- \mid X)} + \frac{F_{\theta_T}(t \mid X)^2 \mathbb{1}\left[U > t\right]}{\overline{G}(t \mid X)}\right]. \tag{3}$$

We derive equality with Equation (2) in Appendix D. Negative Bernoulli log likelihood (BLL) is similar, but with log loss in place of squared error (Appendix E). Viewing the categorical model as a Bernoulli model, one can also quantity model uncertainty at each t independently of the number of categorical bins using e.g. Bernoulli entropy. BS and BLL are strictly proper for classification at a time t, so their summed variants over all times are strictly proper for discrete survival distributions.

Proper objectives differ. Though negative log likelihood (NLL), BS and BLL all have the same true distribution at optimum with infinite data, they may yield significantly different solutions in practice. For example, likelihood incurs infinite loss when a datapoint is assigned 0 probability while BS does not, meaning BS's finite data solutions may be less over-dispersed [Zhang et al., 2021]. Because of this, one should not rely on one loss to yield good performance under another. When a practitioner requires good performance under BS or BLL, they should optimize directly for those metrics.

Re-weighting dilemma. Censoring introduces challenges because we must use IPCW. Crucially, the G in Equation (3) is the true censoring distribution rather than a model, but during training, we only have access to models. This poses a dilemma: can the models be used in re-weighting estimates during training to successfully optimize the criteria as it would be computed without censoring?

4 Inverse-Weighted Survival Games

As a first attempt to solve the dilemma, consider this optimization: choose a failure loss re-weighted by the censoring model, a censoring loss re-weighted by the failure model, and jointly optimize their sum. The expectation is that both models will improve over training and yield reliable IPCW estimates for each other. Concretely, consider optimizing Equation (3) plus the same objective with the roles of F_θ and G_θ reversed. Unfortunately, there exist solutions with smaller loss than the true pair of distributions, making the summed objective improper for the pair of distributions. In Figure 1, we plot this phenomenon for IPCW $\text{BS}(t = 1)$ for models over two timesteps.[2]

To address this phenomenon, we introduce *Inverse-Weighted Survival Games*. These games simultaneously optimize a failure model and censoring model which are non-adversarially featured in each other's losses. We show in experiments that these games produce models with good BS, BLL, and concordance relative to those trained with maximum likelihood.

4.1 Constructing the games

We present and analyze the games for marginal categorical models. The results can be extended to conditional parameterizations with the usual caveats shared by maximum likelihood. Our experiments explore the conditional setting. We specify an Inverse-Weighted Survival Game as follows:

[2] $\text{BS}(t = 1)$ is proper for distributions with support over two timesteps because $\text{BS}(t = K)$ for a model with support over K steps is always 0, so the summed BS equals $\text{BS}(1)$.

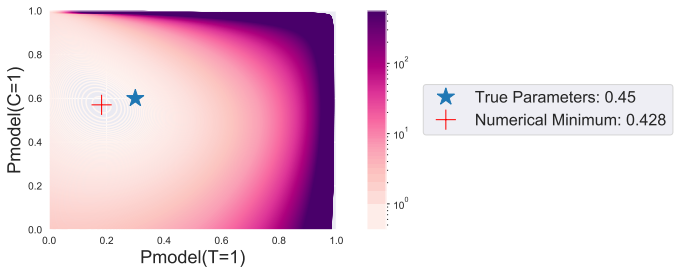


Figure 1: Sum of IPCW brier scores is an improper joint objective for failure and censoring models

First, choose a loss L used for both models. Next, derive the IPCW form L_I that can be used to compute L under censoring. [3] In the game, each model optimizes its loss while serving as the re-weighting distribution in the other model's loss:

$$V_F(\theta) = L_I(F_{\theta_T}; G_{\theta_C}), \quad V_G(\theta) = L_I(G_{\theta_C}; F_{\theta_T}) \tag{4}$$

Compared to Equation (3), we have replaced the true re-weighting distributions with models. The *failure player* minimizes V_F w.r.t. θ_T and the *censor player* minimizes V_G w.r.t. θ_C. One instantiation of these games is the IPCW BS(t) game, derived in Appendix D. With $\overline{\Delta} = 1 - \Delta$,

$$\begin{aligned} V_F^t(\theta) &= \mathbb{E}\left[\frac{\overline{F}_{\theta_T}(t)^2 \Delta \mathbb{1}\left[U \leq t\right]}{\overline{G}_{\theta_C}(U^-)} + \frac{F_{\theta_T}(t)^2 \mathbb{1}\left[U > t\right]}{\overline{G}_{\theta_C}(t)}\right] \\ V_G^t(\theta) &= \mathbb{E}\left[\frac{\overline{G}_{\theta_C}(t)^2 \overline{\Delta} \mathbb{1}\left[U \leq t\right]}{\overline{F}_{\theta_T}(U)} + \frac{G_{\theta_C}(t)^2 \mathbb{1}\left[U > t\right]}{\overline{F}_{\theta_T}(t)}\right]. \end{aligned} \tag{5}$$

As we show next, this formulation has formal advantages over the optimization in Figure 1 for particular choices of L.

Multiple Timesteps. The example is specified for a given t, but categorical BS and BLL are only proper when the whole set from $t = 1$ to $t = K - 1$ is minimized. Two game algorithms can build survival models: solving the games for all timesteps simultaneously with respect to one pair of models (F_θ, G_θ), or playing a single game with value functions equal to the sums $V_F = \sum_t V_F^t$, $V_G = \sum_t V_G^t$. We study the summed games empirically, but prove properties about both approaches.

4.2 IPCW games have a stationary point at data distributions

Among a game's stationary points should be the true failure and censoring distributions.

Proposition 1. Assume $\exists \theta_T^\star \in \Theta_T, \exists \theta_C^\star \in \Theta_C$ such that $F^\star = F_{\theta_T^\star}$ and $G^\star = G_{\theta_C^\star}$. Assume L is proper. Then $(\theta_T^\star, \theta_C^\star)$ is a stationary point of the game Equation (4).

The proof is in Appendix I. The result holds for both the summed and all-timesteps games. BS(t), BLL(t), and AUC(t) and their summed variants are examples of proper scoring rules. [4] When games are built from such objectives, the set of solutions includes $(\theta_T^\star, \theta_C^\star)$, meaning the models do not leave the correct failure and censoring distributions once reached. This result also holds for continuous distributions when using proper rules such as the integrated BS.

4.3 Uniqueness for Discrete Brier Games

We now provide a stronger result for the IPCW BS game in Equation (5) for the all-timesteps variant of the game: its only stationary point is located at the true failure and censoring distributions. This means that IPCW BS games do not introduce extra local optima beyond those expected when introducing deep conditional parameterizations.

[3] For true failure and censoring distributions $F^\star, G^\star$, the losses L and L_I are related through $L_I(F_{\theta_T}; G^\star) = L(F_{\theta_T})$ and $L_I(G_{\theta_C}; F^\star) = L(G_{\theta_C})$.

[4] Interestingly, though often reported, the time-dependent concordance(t) is not proper [Blanche et al., 2019].

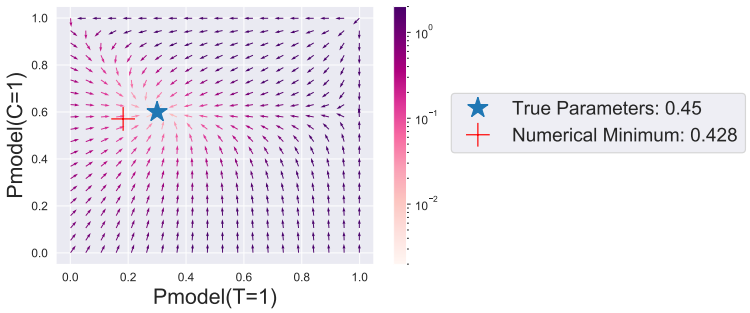


Figure 2: Gradient field of game at timestep one. The unique stationary point of game is at true distribution, contrasting the improper objective in Figure 1.

Proposition 2. Assuming that $\theta^\star_{Tt} > 0$ and $\theta^\star_{Ct} > 0$, the solution $(\theta^\star_T, \theta^\star_C)$ is the only feasible stationary point common to the set games at all times.

The proof is in Appendix J. To illustrate this, Figure 2 shows that, for the same experiment as in Figure 1, the IPCW BS game moves to the correct solution at its unique stationary point.

5 Experiments

We run experiments on a simulation with conditionally Gamma times, a semi-simulated survival dataset based on MNIST, several sources of cancer data, and data on critically-ill hospital patients.

Losses. We build categorical models in 3 ways: the standard NLL method (Equation (6)), the IPCW BS game and the negative IPCW BLL game.

Metrics. For these models we report BS (uncensored for simulations and Kaplan-Meier (KM)-weighted for real data), BLL (also uncensored or weighted), concordance which measures the proportion of pairs whose predicted risks are ordered correctly [Harrell Jr et al., 1996], and NLL.

Model Description. In all experiments except for MNIST, we use a 3-hidden-layer ReLU network that outputs 20 categorical bins (more bin choices in Appendix H.1). For MNIST we first use a small convolutional network and follow with the same fully-connected network (details in Appendix G).

5.1 Simulation Studies

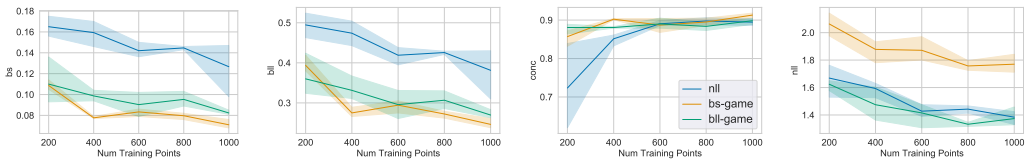
Data. We draw $X \in \mathbb{R}^{32} \sim \mathcal{N}(0, 10I)$ and $T \sim \text{Gamma}(\text{mean} = \mu_t)$ where μ_t a log-linear function of X. The censoring times are also gamma with mean $0.9 * \mu_t$. Both distributions have constant variance 0.05. It holds that $T \perp\!\!\!\perp C \mid X$. Each random seed draws a new dataset.

Results. Figure 3 demonstrates that the games optimize the true uncensored BS, and, though more slowly with respect to training size, log-likelihood does too. The games have better test-set performance on all metrics for small training size. All methods converge on similar performance when there is enough data (though *enough* is highly-dependent on dimensionality and model class).

Calibration. We include a qualitative comparison investigating model calibration on the gamma simulation trained with 2000 datapoints. Figure 4 shows that the BS game achieves near-perfect calibration while the two likelihood-based methods suffer some error. This is expected since likelihood-based methods do not directly optimize calibration while BS does (Section 3).

5.2 Semi-simulated studies

Data. Survival-MNIST [Gensheimer, 2019, Pölsterl, 2019] draws times conditionally on MNIST label Y. This means digits define risk groups and $T \perp\!\!\!\perp X \mid Y$. Times within a digit are i.i.d. The model only sees the image pixels X as covariates so it must learn to classify digits (risk groups) to model



(a) Uncensored BS (b) Uncensored Negative BLL (c) Concordance (d) Categorical NLL

Figure 3: Test set evaluation metrics on Gamma simulation versus number of training points for three different methods. Lower is better for all the metrics except for concordance. text larger

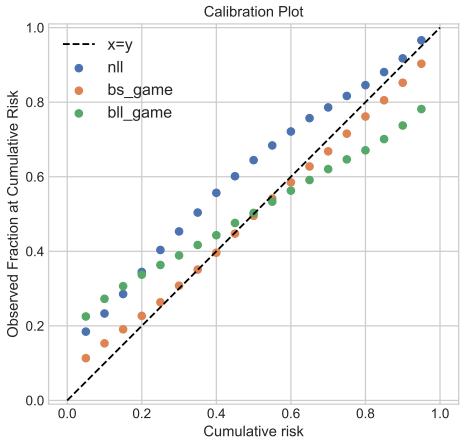
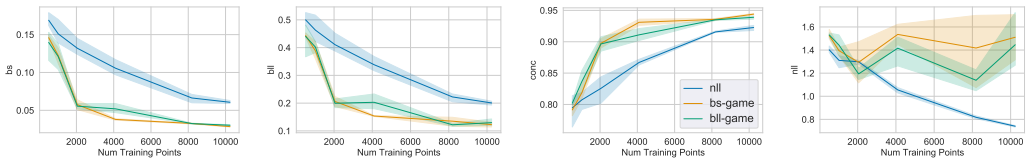


Figure 4: todo This is a caption placeholder for a caption that will be two lines. Should the caption explain why this plot shows calibration? .

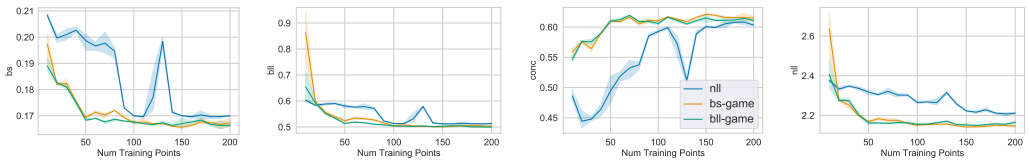
times. We follow Goldstein et al. [2020] and use Gamma times. $T \sim \text{Gamma}(\text{mean} = 10 * (Y + 1))$. We set the variance constant to 0.05. Lower digit labels Y yield earlier event times. C is drawn similarly but with mean $9.9 * (Y + 1)$. Each random seed draws a new dataset.

Results. This experiment demonstrates that better NLL does not correspond to better performance on BS, BLL, and concordance. Similarly to the previous experiment, Figure 5 shows that game methods attain better uncensored test-set BS and BLL on survival-MNIST than likelihood-based training does. The games likewise attain higher concordance. NLL training performs better at the metric it directly optimizes. This experiment also establishes that it is possible to optimize through deep convolutional models with batch norm and pooling using the game training methods.



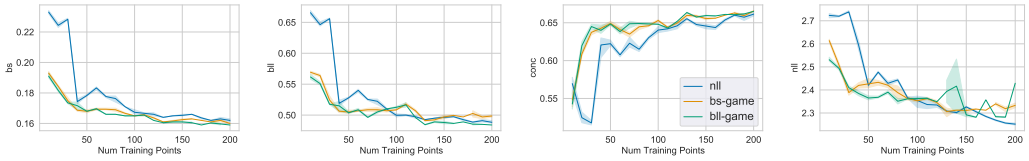
(a) Uncensored BS (b) Uncensored Negative BLL (c) Concordance (d) Categorical NLL

Figure 5: Test set evaluation metrics on the survival-MNIST simulation versus number of training points for three methods. Lower is better for all the metrics except for concordance. text larger



(a) KM-weighted BS (b) KM-weighted Negative BLL (c) concordance (d) categorical NLL

Figure 6: Test set evaluation metrics on METABRIC versus number of training points for three methods. Lower is better for all the metrics except for concordance. text larger



(a) KM-weighted BS (b) KM-weighted Negative BLL (c) concordance (d) categorical NLL

Figure 7: Test set evaluation metrics on ROTT. & GBSG versus number of training points for three methods. Lower is better for all the metrics except for concordance. text larger

5.3 Real Datasets

Data. We use several datasets used in recent papers [Chen, 2020, Kvamme et al., 2019] and available in the python packages DeepSurv [Katzman et al., 2018] and PyCox [Kvamme et al., 2019], and the R Survival [Therneau, 2021]. The datasets are:

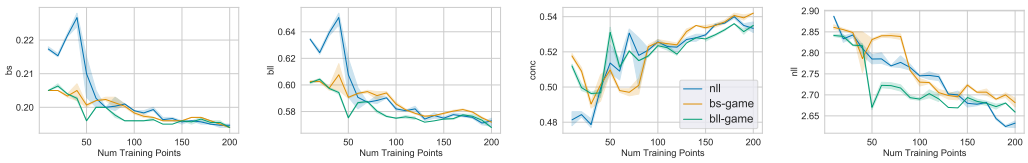
- Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [Curtis et al., 2012]
- Rotterdam Tumor Bank (ROTT) [Foekens et al., 2000] and German Breast Cancer Study Group (GBSG) [Schumacher et al., 1994] combined into one dataset (ROTT. & GBSG)
- Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT) [Knaus et al., 1995] which includes severely ill hospital patients

For more description see Appendix G. For real data, there is no known ground truth for the censoring distribution, which means evaluation requires assumptions. Following the experiments in Kvamme et al. [2019], we assume that censoring is marginal estimate with KM to evaluate models. [5]

Results. On METABRIC, games attain lower (better) KM-weighted BS and BLL than NLL-training when the number of datapoints is small, and have better concordance and NLL though they do not directly optimize them. As data size increases, all methods converge to similar performance. On ROTT. & GBSG, the trend is similar: games optimize the BS and BLL more rapidly as a function of training set size than NLL-training does. Again, all methods converge to similar performance in all metrics when the number of datapoints is large enough. All methods perform similarly on SUPPORT.

Caveats. First, though popular, these survival datasets are low-dimensional, so any of the objectives can perform well on the metrics with just several hundred points. We see that this is distinct from MNIST, where thousands of points were required to improve performance. Second, though possible, it may not be true that censoring is marginal on these datasets, which would mean that the BS and BLL results only have their interpretation conditional on a particular set of covariates. Lastly, no method is stable for all metrics, for all training sizes, on all seeds for all datasets.

[5] This is also the route taken in the R packages Survival [Therneau, 2021], PEC [Mogensen et al., 2012], and riskRegression [Gerds et al., 2020].



(a) KM-weighted Negative BLL (b) concordance (c) categorical NLL

Figure 8: Test set evaluation metrics on SUPPORT versus number of training points for three methods. Lower is better for all the metrics except for concordance. text larger

6 Related Work

Nuisance parameters. Under non-informative censoring, the censoring distribution is unrelated to the failure distribution, but estimating it can help improve learning the failure distribution; here, the censoring distribution is a *nuisance parameter*. Existing causal estimation methods propose two-stage procedures where the first stage estimates the nuisance-parameter and the second stage uses the learned nuisance-parameter as-is to define an estimator or loss function for the target parameter. [Van Der Laan and Rubin, 2006, Van der Laan and Rose, 2011, Chernozhukov et al., 2018, Foster and Syrgkanis, 2019]. In this work, we instead show that estimating the target (failure model objective or failure model itself) can benefit from a *coupled* estimation procedure where the nuisance parameter (censoring model) is also trained simuteneously. The failure model needs the censoring distribution to compute BS but censoring estimation needs the failure model, and despite this circular dependence, we characterize a case where the optimization leads to the two true distributions.

Double Robust Censoring Unbiased Transformations. For functions h, Rubin and van der Laan [2007] estimate conditional mean $\mathbb{E}[h(T, X)|X]$ under censoring using a double-robust estimator: given estimates of the conditional failure and censoring CDFs $\hat{F}(t|X)$ and $\hat{G}(c|X)$, the estimator of $\mathbb{E}[h(T, X)|X]$ is unbiased when either nuisance CDF is correct. However, here we are concerned with estimating a quantity to be used as a loss for learning $\hat{F}$. We therefore presumably do not already have an estimate of $\hat{F}$ to be used in a doubly-robust estimator.

Censoring Unbiased Losses for Deep Learning. Steingrimsson and Morrison [2020] build loss functions for the failure model from the estimators in Rubin and van der Laan [2007]. In particular, their BS loss extends IPCW BS estimation to the doubly-robust case and to our knowledge is the first instance of IPCW-based estimation procedures being used in a general purpose way to define loss functions for deep survival analysis. There however, the censoring distribution is estimated once before training and held fixed rather than incorporated into a joint training procedure as in the games in this work. The fixed censoring estimate is implemented by KM, which assumes a marginal censoring distribution. Making the marginal assumption when censoring is truly conditional should not yield a performant model under the BS criteria since the training objective does not directly estimate or optimize the true BS that would be measured under no censoring. When marginal censoring does hold, KM estimation, which is non-parametric, may be a simpler and stable choice versus the game, depending on sample size, data variance, and conditional parameterization assumptions. But since it is in general unknown if censoring is marginal, we use conditional models.

7 Discussion

In this work, we propose a new training method for survival models under censored data. We argue that on finite data, it is important to close the gap between training methodology and the desired evaluation criteria. We showed in the experiments that better NLL does not correspond to better performance on BS, BLL, and concordance, all evaluations of interest in survival analysis.

The main trend in our experimental results was that data size matters: smaller meant the game methods performed better than NLL and enough data meant that they perform similarly, which is expected since all objectives are proper. However *enough* data is hard to define: it depends on

dimensionality and on the data generating distribution and model class. It is a great direction of future work to build more precise understanding on how objectives behave differently even when they have the same optimum on infinite data.

Limitations. As mentioned, evaluation on real data under censoring requires assumptions. It is important to further consider how to better assess test-set performance on metrics such as BS, BLL, and concordance. Because concordance is not proper, we do not build objectives from it here, but it too is not invariant to censoring. Regarding games, we showed properties about statonary points. More analysis is necessary to describe important convergence properties of optimizing these games.

Future Work. Should we include this?

Social Impact. Survival models are deployed in hospital settings and have high impact on public health. In this work, we saw benefits of a new training approach for these models, but no training method is a panacea. Practitioners of survival analysis must take great care to consider various training and validation approaches, as well as consider possible test distribution shifts, prior to deployment.

Acknowledgments and Disclosure of Funding

List funding, competing interests, people that helped.

References

A. M. Alaa and M. van der Schaar. Deep multi-task gaussian processes for survival analysis with competing risks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2326–2334. Curran Associates Inc., 2017.

P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

P. Blanche, J.-F. Dartigues, and H. Jacqmin-Gadda. Review and comparison of roc curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55(5): 687–704, 2013.

P. Blanche, M. W. Kattan, and T. A. Gerds. The c-index is not proper for the evaluation of-year predicted risks. *Biostatistics*, 20(2):347–357, 2019.

G. W. Brier and R. A. Allen. Verification of weather forecasts. In *Compendium of meteorology*, pages 841–848. Springer, 1951.

G. H. Chen. Deep kernel survival analysis and subject-specific survival time prediction intervals. In *Machine Learning for Healthcare Conference*, pages 537–565. PMLR, 2020.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

S. R. Cole and M. A. Hernán. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664, 2008.

C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

O. V. Demler, N. P. Paynter, and N. R. Cook. Tests of calibration and goodness-of-fit in the survival setting. *Statistics in medicine*, 34(10):1659–1680, 2015.

M. R. Elliott. Model averaging methods for weight trimming. *Journal of official statistics*, 24(4):517, 2008.

J. A. Foekens, H. A. Peters, M. P. Look, H. Portengen, M. Schmitt, M. D. Kramer, N. Brünner, F. Jänicke, M. E. Meijer-van Gelder, S. C. Henzen-Logmans, et al. The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer research*, 60(3): 636–643, 2000.

D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

S. Fotso. Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*, 2018.

B. Gensheimer, Michael F.and Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ 7:e6257*, 2019.

T. A. Gerds and M. Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.

T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32 (13):2173–2184, 2013.

T. A. Gerds, P. Blanche, T. H. Scheike, R. Mortensen, M. Wright, N. Tollenaar, J. Muschelli, U. B. Mogensen, and B. Ozenne. Package 'riskregression', 2020.

G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

M. Goldstein, X. Han, A. M. Puli, A. Perotte, and R. Ranganath. X-cal: Explicit calibration for survival analysis. *Advances in Neural Information Processing Systems*, 33, 2020.

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

H. Haider, B. Hoehn, S. Davis, and R. Greiner. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85):1–63, 2020.

P. T. Harker and J.-S. Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 48(1-3): 161–220, 1990.

F. E. Harrell Jr, K. L. Lee, and D. B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

X. Huang, R. A. Wolfe, and C. Hu. A test for informative censoring in clustered survival data. *Statistics in medicine*, 23(13):2089–2107, 2004.

H. Hung and C.-T. Chiang. Estimation methods for time-dependent auc models with survival data. *Canadian Journal of Statistics*, 38(1):8–26, 2010a.

H. Hung and C.-t. Chiang. Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian journal of statistics*, 37(4):664–679, 2010b.

E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2 edition, 2002.

F. Kamran and J. Wiens. Estimating calibrated individualized survival curves with deep learning. 2021.

J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

W. A. Knaus, F. E. Harrell, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors, N. V. Dawson, W. J. Fulkerson, R. M. Califf, N. Desbiens, et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, 122(3):191–203, 1995.

A. Kumar, S. Sarawagi, and U. Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pages 2805–2814, 2018.

H. Kvamme and Ø. Borgan. The brier score under administrative censoring: Problems and solutions. *arXiv preprint arXiv:1912.08581*, 2019a.

H. Kvamme and Ø. Borgan. Continuous and discrete-time survival prediction with neural networks. *arXiv preprint arXiv:1910.06724*, 2019b.

H. Kvamme, Ørnulf Borgan, and I. Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019. URL http://jmlr.org/papers/v20/18-424.html.

B. K. Lee, J. Lessler, and E. A. Stuart. Weight trimming and propensity score weighting. *PloS one*, 6 (3):e18174, 2011.

C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

C. Lee, W. Zame, A. Alaa, and M. Schaar. Temporal quilting for survival analysis. In *The 22nd international conference on artificial intelligence and statistics*, pages 596–605. PMLR, 2019.

X. Miscouridou, A. Perotte, N. Elhadad, and R. Ranganath. Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, pages 244–256, 2018.

U. B. Mogensen, H. Ishwaran, and T. A. Gerds. Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software*, 50(11):1–23, 2012. URL https://www.jstatsoft.org/v50/i11.

S. Pölsterl. Sebastian pölsterl, Jul 2019. URL https://k-d-w.org/blog/2019/07/survival-analysis-for-deep-learning/.

R. Ranganath, A. Perotte, N. Elhadad, and D. Blei. Deep survival analysis. *arXiv preprint arXiv:1608.02158*, 2016.

K. Ren, J. Qin, L. Zheng, Z. Yang, W. Zhang, L. Qiu, and Y. Yu. Deep recurrent survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4798–4805, 2019.

D. Rubin and M. J. van der Laan. A doubly robust censoring unbiased transformation. *The international journal of biostatistics*, 3(1), 2007.

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448): 1096–1120, 1999.

M. Schumacher, G. Bastert, H. Bojar, K. Hübner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R. Neumann, and H. Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093, 1994.

M. Sloma, F. Syed, M. Nemati, and K. S. Xu. Empirical comparison of continuous and discrete-time representations for survival prediction. In *Survival Prediction-Algorithms, Challenges and Applications*, pages 118–131. PMLR, 2021.

J. A. Steingrimsson and S. Morrison. Deep learning for survival outcomes. *Statistics in medicine*, 39 (17):2339–2349, 2020.

L. M. Sullivan, J. M. Massaro, and R. B. D'Agostino Sr. Presentation of multivariate data for clinical use: The framingham study risk score functions. *Statistics in medicine*, 23(10):1631–1660, 2004.

T. M. Therneau. *A Package for Survival Analysis in R*, 2021. URL https://CRAN.R-project.org/package=survival. R package version 3.2-11.

H. Uno, T. Cai, L. Tian, and L.-J. Wei. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537, 2007.

M. J. Van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

M. J. Van Der Laan and D. Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

M. J. Van der Laan, M. Laan, and J. M. Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016.

M. Wolbers, P. Blanche, M. T. Koller, J. C. Witteman, and T. A. Gerds. Concordance for prognostic models with competing risks. *Biostatistics*, 15(3):526–539, 2014.

S. Yadlowsky, S. Basu, and L. Tian. A calibration metric for risk scores with survival data. In *Machine Learning for Healthcare Conference*, pages 424–450, 2019.

C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in Neural Information Processing Systems*, pages 1845–1853, 2011.

L. Zhang, M. Goldstein, and R. Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *International Conference on Machine Learning*, pages 12427–12436. PMLR, 2021.

C. Zhong and R. Tibshirani. Survival analysis as a classification problem. *arXiv preprint arXiv:1909.11171*, 2019.

A TODO

need to cite scrps somewhere (see latex comments below) and possibly include it in plots?

todo revise plots. scatter instead of line and make captions clear.

todo where to put this While IPCW BS and IPCW BLL are re-weighted to be invariant to the censoring distribution, NLL's scale is censoring distribution-dependent (Appendix B.4).

check if experiment description in main text and in appendix matches gamma generation in code for final results.

clean up game algorithm appendix section to match summed vs all-timesteps description in main text the training is also okay under marginally independent censoring T indep C. We have updated the text to mention this. where to mention this?

when does a wrongly weighted bs still optimize the true bs at optimum? For example if you marginally-KM-weighted on a conditional censoring task or if you use unweighted BS for a marginally or conditionally censored task. We know these wont unbiased estimate the true bs, but we wonder when it can still optimize it. This affects our discussion about Steingrimmson and about SCRPS. km weighted ablation from rebuttal in appendix?

B Notation, Assumptions, and Likelihoods in More Detail

B.1 Notation

Let T be a failure time with CDF F. T's *survival function* is defined by $\overline{F} = 1 - F$. We denote failure models by F_{θ_T}. Let C be a censoring time with CDF G, survival function $\overline{G}$, and model G_{θ_C}. Under right-censoring, $U = \min(T, C)$, $\Delta = \mathbb{1}[T \leq C]$ and we observe (X_i, U_i, Δ_i). We use $\overline{G}(t^-)$ to denote $P(C \geq t)$.

B.2 Assumptions

We assume i.i.d. data and random censoring: $T \perp\!\!\!\perp C \mid X$ [Kalbfleisch and Prentice, 2002]. Derivations in this work also require the censoring positivity assumption [Gerds et al., 2013]. Let f be a failure density or mass,

$$\forall x, \forall t \in \{t \mid f(t|x) > 0\}, \quad \exists \epsilon \quad \text{s.t.} \quad \overline{G}(t^-|x) > \epsilon > 0.$$

This says we observe all subjects' events with positive probability. "Censoring" typically refers to missingness of failure times. However, in this work we model censoring too, a survival modeling task where censoring times are censored by failures. We then also need the positivity condition to hold with the roles of F and G reversed in order to observe censoring events properly. Let g be a censoring density or mass,

$$\forall x, \forall t \in \{t \mid g(t|x) > 0\}, \quad \exists \epsilon \quad \text{s.t.} \quad \overline{F}(t|x) > \epsilon > 0.$$

B.3 Likelihoods

As mentioned, we assume data are i.i.d. and censoring is random $T \perp\!\!\!\perp C \mid X$. Under these assumptions, the likelihood, by definition [Andersen et al., 2012], is:

$$L(\theta_t, \theta_c) = \prod_i \left[f_{\theta_t}(U_i) \overline{G}_{\theta_c}(U_i^-) \right]^{\Delta_i} \left[g_{\theta_c}(U_i) \overline{F}_{\theta_t}(U_i) \right]^{1-\Delta_i}, \tag{6}$$

When a failure is observed, $\Delta_i = \mathbb{1}[T_i \leq C_i] = 1$ so we compute the failure density or mass f at the observed time $U_i = T_i$. In this case, the only thing we know about the censoring time is $C_i \geq T_i = U_i$. We therefore compute $P(C_i \geq T_i) = P(C_i \geq U_i) = 1 - G_{\theta_c}(U_i^-) = \overline{G}_{\theta_c}(U_i^-)$. Likewise, when a censoring time is observed, $\Delta_i = 0$ so we compute the censoring density or mass g at the observed censoring time $U_i = C_i$. In this case, the only thing we know about the failure time is that $T_i > C_i$. We therefore compute $P(T_i > C_i) = P(T_i > U_i) = 1 - F(U_i) = \overline{F}(U_i)$.

Under the additional assumption of non-informativeness -that F and G don't share parameters and therefore θ_t, θ_c are distinct- the g/G terms are constant wrt θ_t and the f/F terms are constant wrt θ_c. In this case, when one is modeling failures, they can use the partial failure likelihood:

$$L(\theta_t)^{\text{partial}} = \prod_i \left[f_{\theta_t}(U_i)\right]^{\Delta_i} \left[\overline{F}_{\theta_t}(U_i)\right]^{1-\Delta_i}$$

And when one is modeling censoring they can use the partial censoring likelihood:

$$L(\theta_c)^{\text{partial}} = \prod_i \left[\overline{G}_{\theta_t}(U_i^-)\right]^{\Delta_i} \left[g_{\theta_t}(U_i)\right]^{1-\Delta_i}$$

B.4 failure partial likelihood depends on censoring

update this to match paper notation rather than rebuttal notation

we now show that the failure partial likelihood's scale depends on the true sampling distribution of censoring times, even if the censoring model has dropped as a constant in the objective

$$E_{T\sim F_{true}, C\sim G_{true}}[p_\theta(U)^{\Delta=1}\overline{F}_\theta(U)^{\Delta=0}]$$

Here, Δ and U depend on T and C (therefore on F_{true} and G_{true}). We now constructively show that the failure model's NLL can vary with the true censoring distribution. Let us consider a marginal survival analysis problem (no features) and random censoring. The log NLL is:

$E_{F_{true},G_{true}}[\Delta \log p_\theta(U)] + E_{Ftrue,Gtrue}[(1-\Delta)\log \overline{F}_\theta(U)]$

Now consider an F_{true} whose support starts at time 1 (e.g. uniform over 1,2,3) and G_{true} such that there is probability ρ that $C = 0$ and probability $1-\rho$ that C take a value above the support of T (e.g. >3). Points are therefore only censored at time 0 or uncensored.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{F_{true},G_{true}}[\Delta \log p_\theta(U)] + \mathop{\mathbb{E}}_{F_{true},G_{true}}[(1-\Delta)\log \overline{F}_\theta(U)] \\
&= (1-\rho) \mathop{\mathbb{E}}_{F_{true}}[\log p_\theta(T)] + \rho \mathop{\mathbb{E}}_{G_{true}}[\log \overline{F}_\theta(C)] \\
&= (1-\rho) \mathop{\mathbb{E}}_{F_{true}}[\log p_\theta(T)] + \rho \mathop{\mathbb{E}}_{G_{true}}[\log \overline{F}_\theta(0)] \\
&= (1-\rho) \mathop{\mathbb{E}}_{F_{true}}[\log p_\theta(T)] + \rho \mathop{\mathbb{E}}_{G_{true}}[\log 1] \\
&= (1-\rho) \mathop{\mathbb{E}}_{F_{true}}[\log p_\theta(T)] + \rho \mathop{\mathbb{E}}_{G_{true}}[0] \\
&= (1-\rho) \mathop{\mathbb{E}}_{F_{true}}[\log p_\theta(T)]
\end{aligned}$$

This quantity depends on ρ. This shows that the failure model's NLL depends on the true sampling distribution of censoring times.

C IPCW Primer

IPCW is a technique for estimation under censoring [Gerds and Schumacher, 2006]. Consider estimating the marginal mean of T : $\mathbb{E}[T] = \mathbb{E}_X \mathbb{E}_{T|X}[T]$. T is not observed for all datapoints. Instead, we observe $U = \min(T, C)$ and $\Delta = \mathbb{1}[T \leq C]$. IPCW reformulates such expectations in terms of observed data. Using this method, we can show that:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{X,T}[T] &= \mathop{\mathbb{E}}_{X} \mathop{\mathbb{E}}_{T|X} \left[\frac{\mathbb{E}_{C|X} \mathbb{1}[T \leq C]}{\mathbb{E}_{C|X} \mathbb{1}[T \leq C]} T \right] \\
&= \mathop{\mathbb{E}}_{X} \mathop{\mathbb{E}}_{T|X} \mathop{\mathbb{E}}_{C|X} \left[\frac{\mathbb{1}[T \leq C]}{\mathbb{E}_{C'|X} \mathbb{1}[T \leq C']} T \right] \\
&= \mathop{\mathbb{E}}_{T,C,X} \left[\frac{\mathbb{1}[T \leq C]}{\mathbb{E}_{C'|X} \mathbb{1}[T \leq C']} T \right] \\
&= \mathop{\mathbb{E}}_{T,C,X} \left[\frac{\mathbb{1}[T \leq C]}{\mathbb{P}(C' \geq T|X)} T \right] \\
&= \mathop{\mathbb{E}}_{T,C,X} \left[\frac{\mathbb{1}[T \leq C]}{\overline{G}(T^-|X)} T \right] \\
&= \mathop{\mathbb{E}}_{U,\Delta,X} \left[\frac{\Delta U}{\overline{G}(U^-|X)} \right]
\end{aligned}
$$

We have used C' in the denominator to emphasize that it is not a function of C in the integral over the numerator indicator once that expectation is moved out. We have used random censoring to go from $\mathbb{E}_{T|X}\, \mathbb{E}_{C|X}$ to the joint $\mathbb{E}_{T,C|X}$. The last equality changes from the complete data distribution to the observed distribution and holds because $\Delta = 1 \implies U = T$. This means we can estimate the expectation, provided that we know G and that random censoring and positivity (Equation (1)) hold. In practice, we must learn the censoring distribution, a challenging task as it is also censored.

Graf et al. [1999] develop the IPCW BS. Gerds and Schumacher [2006] extend it to conditional censoring and Kvamme and Borgan [2019a] specialize to administrative censoring. Gerds et al. [2013], Wolbers et al. [2014] develop the IPCW concordance. IPCW estimators for several forms of area under curve (AUC) have been studied in Hung and Chiang [2010a,b], Blanche et al. [2013, 2019], Uno et al. [2007]. Yadlowsky et al. [2019] derive an IPCW estimator for binary survival calibration.

Even under positivity, $\overline{G}(t)$ can become so small that the weights blow up, causing enormous variance. In practice, clamping $\overline{G}(t)$ to be in e.g. $[0.05, 1]$ is used. Further discussion can be found in the literature on truncated (or "clipped") importance sampling [Ionides, 2008], propensity weighting [Elliott, 2008, Cole and Hernán, 2008, Scharfstein et al., 1999, Lee et al., 2011] and off-policy evaluation in reinforcement learning (e.g. [Wang et al., 2016]).

D Deriving IPCW Brier Scores

We derive the IPCW BS introduced by Graf et al. [1999], Gerds and Schumacher [2006]. The censor-weighted failure BS:

$$\text{F-BS-CW}(t) = \mathop{\mathbb{E}}_{T,C}\left[\frac{(1 - F_\theta(t))^2 \mathbb{1}\left[T \leq C\right] \mathbb{1}\left[U \leq t\right]}{P_\theta(C' \geq U)} + \frac{F_\theta(t)^2 \mathbb{1}\left[U > t\right]}{P_\theta(C' > t)}\right]$$

where $U = \min(T, C)$ and $F_\theta = P_\theta(T \leq \cdot)$, It's relationship to the regular BS is:

$$\begin{aligned}
\text{F-BS}(t) &= \mathop{\mathbb{E}}_{T}\left[\Big(F_\theta(t) - \mathbb{1}\left[T \leq t\right]\Big)^2\right] \\
&= \mathop{\mathbb{E}}_{T}\left[(1 - F_\theta(t))^2 \mathbb{1}\left[T \leq t\right] + F_\theta(t)^2 \mathbb{1}\left[T > t\right]\right] \\
&= \mathop{\mathbb{E}}_{T}\left[\frac{\mathbb{E}_C \mathbb{1}\left[T \leq C\right]}{\mathbb{E}_C \mathbb{1}\left[T \leq C\right]}(1 - F_\theta(t))^2 \mathbb{1}\left[T \leq t\right] + \frac{\mathbb{E}_C \mathbb{1}\left[C > t\right]}{\mathbb{E}_C \mathbb{1}\left[C > t\right]} F_\theta(t)^2 \mathbb{1}\left[T > t\right]\right] \\
&= \mathop{\mathbb{E}}_{T,C}\left[\frac{(1 - F_\theta(t))^2 \mathbb{1}\left[T \leq C\right] \mathbb{1}\left[T \leq t\right]}{\mathbb{E}_C \mathbb{1}\left[T \leq C\right]} + \frac{F_\theta(t)^2 \mathbb{1}\left[T > t\right] \mathbb{1}\left[C > t\right]}{\mathbb{E}_C \mathbb{1}\left[C > t\right]}\right] \\
&= \mathop{\mathbb{E}}_{T,C}\left[\frac{(1 - F_\theta(t))^2 \mathbb{1}\left[T \leq C\right] \mathbb{1}\left[T \leq t\right]}{P_\theta(C' \geq T)} + \frac{F_\theta(t)^2 \mathbb{1}\left[T > t\right] \mathbb{1}\left[C > t\right]}{P_\theta(C' > t)}\right] \\
&= \mathop{\mathbb{E}}_{T,C}\left[\frac{(1 - F_\theta(t))^2 \mathbb{1}\left[T \leq C\right] \mathbb{1}\left[U \leq t\right]}{P_\theta(C' \geq U)} + \frac{F_\theta(t)^2 \mathbb{1}\left[U > t\right]}{P_\theta(C' > t)}\right] \\
&= \text{F-BS-CW}(t)
\end{aligned}$$

The expectation comes out due to $T \perp\!\!\!\perp C$. The last line follows from $T \leq C \implies U = T$ (in the left term) and $\mathbb{1}\left[T > t\right] \mathbb{1}\left[C > t\right] = \mathbb{1}\left[U > t\right]$ (in the right term). Define likewise the failure-weighted censor BS

$$\text{G-BS-CW}(t) = \mathop{\mathbb{E}}_{T,C}\left[\frac{(1 - G_\theta(t))^2 \mathbb{1}\left[C < T\right] \mathbb{1}\left[U \leq t\right]}{P_\theta(T' > U)} + \frac{G_\theta(t)^2 \mathbb{1}\left[U > t\right]}{P_\theta(T' > t)}\right]$$

where $G_\theta = P_\theta(C \leq \cdot)$. The relationship to the censoring distribution's BS is:

$$\begin{aligned}
\text{G-BS}(t) &= \mathop{\mathbb{E}}_{C}\left[\Big(G_\theta(t) - \mathbb{1}\left[C \leq t\right]\Big)^2\right] \\
&= \mathop{\mathbb{E}}_{C}\left[(1 - G_\theta(t))^2 \mathbb{1}\left[C \leq t\right] + G_\theta(t)^2 \mathbb{1}\left[C > t\right]\right] \\
&= \mathop{\mathbb{E}}_{C}\left[\frac{\mathbb{E}_T \mathbb{1}\left[C < T\right]}{\mathbb{E}_T \mathbb{1}\left[C < T\right]}(1 - G_\theta(t))^2 \mathbb{1}\left[C \leq t\right] + \frac{\mathbb{E}_T \mathbb{1}\left[T > t\right]}{\mathbb{E}_T \mathbb{1}\left[T > t\right]} G_\theta(t)^2 \mathbb{1}\left[C > t\right]\right] \\
&= \mathop{\mathbb{E}}_{T,C}\left[\frac{(1 - G_\theta(t))^2 \mathbb{1}\left[C < T\right] \mathbb{1}\left[C \leq t\right]}{\mathbb{E}_T \mathbb{1}\left[C < T\right]} + \frac{G_\theta(t)^2 \mathbb{1}\left[T > t\right] \mathbb{1}\left[C > t\right]}{\mathbb{E}_T \mathbb{1}\left[T > t\right]}\right] \\
&= \mathop{\mathbb{E}}_{T,C}\left[\frac{(1 - G_\theta(t))^2 \mathbb{1}\left[C < T\right] \mathbb{1}\left[C \leq t\right]}{P_\theta(T' > C)} + \frac{G_\theta(t)^2 \mathbb{1}\left[T > t\right] \mathbb{1}\left[C > t\right]}{P_\theta(T' > t)}\right] \\
&= \mathop{\mathbb{E}}_{T,C}\left[\frac{(1 - G_\theta(t))^2 \mathbb{1}\left[C < T\right] \mathbb{1}\left[U \leq t\right]}{P_\theta(T' > U)} + \frac{G_\theta(t)^2 \mathbb{1}\left[U > t\right]}{P_\theta(T' > t)}\right] \\
&= \text{G-BS-CW}(t)
\end{aligned}$$

The expectation comes out due to $T \perp\!\!\!\perp C$. The last line follows from $C < T \implies U = C$ (in the left term) and $\mathbb{1}\left[T > t\right] \mathbb{1}\left[C > t\right] = \mathbb{1}\left[U > t\right]$ (in the right term).

E Negative Bernoulli Log Likelihood

Negative BLL is similar to BS, but replaces the squared error with negated log loss:

$$\text{NBLL}(t;\theta) = \mathop{\mathbb{E}}_{T,C,X}\Big[-\log(F_{\theta_T}(t \mid X))\mathbb{1}\left[U \leq t\right] - \log(\overline{F}_{\theta_T}(t \mid X))\mathbb{1}\left[U > t\right]\Big]$$

IPCW BLL can likewise be written as [Kvamme et al., 2019]:

$$\text{F-NBLL-CW}(t;\theta) = \mathop{\mathbb{E}}_{T,C,X}\Big[\frac{-\log(F_{\theta_T}(t \mid X))\Delta\mathbb{1}\left[U \leq t\right]}{G(U^- \mid X)} + \frac{-\log(\overline{F}_{\theta_T}(t \mid X))\mathbb{1}\left[U > t\right]}{G(t \mid X)}\Big]$$

F Game Algorithm

THIS SECTION NEEDS A BIG OVERHAUL, DO NOT READ FOR NOW

todo need to first resolve how the two types of games are mentioned in main text then clean this appendix

todo below is the old main text section about this

F.1 Finding Stationary Points

A solution is a *stationary point*, a set of states where no player will change their state. Such points (θ_T, θ_C) are characterized by the variational inequality [Harker and Pang, 1990, Gidel et al., 2018], which says that at (θ_T, θ_C), all directional derivatives are non-negative. To solve all games simultaneously using unconstrained optimization, the parameters must normalize appropriately. Let h be an invertible function mapping from unconstrained parameters $\eta_{T1:K-1}$ to categorical parameters $\theta_{T1:K-1}$. Assembling the failure player derivatives in vector form, we have:

$$dV_F(\theta) = \left[\left.\frac{dV_F^1}{d\theta_{T1}}\right|_\theta, \quad \ldots, \quad \left.\frac{dV_F^t}{d\theta_{Tt}}\right|_\theta, \quad \ldots, \quad \left.\frac{dV_F^{K-1}}{d\theta_{T,K-1}}\right|_\theta \right]$$

To solve the games, we want $dV_F(\theta) = 0$ and $dV_G(\theta) = 0$. Letting J_h be the Jacobian of h, define:

$$g_T(\eta_T) \triangleq \left(J_h|_{\eta_T} \right)^\top dV_F(h(\eta_T)) \tag{7}$$

We can show that setting $g_T(\eta_T) = 0$ sets $dV_F(h(\eta_T)) = 0$. When h is invertible, J_h^{-1} exists when $J_{h^{-1}}$ does. Then, provided we choose an h with differentiable inverse,

$$\left(J_{h^{-1}}|_{h(\eta_T)} \right)^\top g_T(\eta_T) = dV_F(h(\eta_T))$$

An invertible matrix only has the 0-vector in its kernel, which means $g_T(\eta_T) = 0 \iff \forall t, \frac{dV_F^t}{d\theta_t} = 0$. Define $g_C(\eta_C)$ analogously. To set $g_T(\eta_T)$ and $g_C(\eta_C)$ to 0 one can directly minimize their norms. In our experiments, following g_T and g_C converges (Appendix F.1 and Appendix F).

todo below is the old appendix section about this]

We describe the gradient computation. Recall that we have unconstrained parameters η_T (for failure modeling) and η_C (for censoring modeling). We use normalization function h top map to probability parameters $\theta_T = h(\eta_T)$ and $\theta_C = h(\eta_C)$. Our goal is to set $\frac{dV_F^t}{d\theta_{Tt}} = 0$ and $\frac{dV_G^t}{d\theta_{Ct}} = 0$ for all t. Assembling the derivatives for all timesteps in vector form, we have:

$$dV_F(\theta) = \left[\left.\frac{dV_F^1}{d\theta_{T1}}\right|_\theta, \quad \ldots, \quad \left.\frac{dV_F^t}{d\theta_{Tt}}\right|_\theta, \quad \ldots, \quad \left.\frac{dV_F^{K-1}}{d\theta_{T,K-1}}\right|_\theta \right]$$

$$dV_G(\theta) = \left[\left.\frac{dV_G^1}{d\theta_{C1}}\right|_\theta, \quad \ldots, \quad \left.\frac{dV_G^t}{d\theta_{Ct}}\right|_\theta, \quad \ldots, \quad \left.\frac{dV_G^{K-1}}{d\theta_{C,K-1}}\right|_\theta \right]$$

Letting J_h be the Jacobian of h, define:

$$g_T(\eta_T) \triangleq \left(J_h|_{\eta_T} \right)^\top dV_F(h(\eta_T))$$

$$g_C(\eta_C) \triangleq \left(J_h|_{\eta_C} \right)^\top dV_G(h(\eta_C))$$

We showed that setting $g_T(\eta_T) = 0 \implies dV_F(h(\eta_T), h(\eta_C)) = 0$ and $g_C(\eta_C) = 0 \implies dV_G(h(\eta_T), h(\eta_C)) = 0$. To set $g_T(\eta_T)$ and $g_C(\eta_C)$ to 0, one can directly minimize their norms. In our experiments, following g_T and g_C converges:

Algorithm 1 Following Gradients in Inverse-Weighted Games (vector form)

Input: Choice of value functions V_F, V_G, normalization function h, learning rate γ.
Initialize $\eta_{T,1:K-1}$ and $\eta_{C,1:K-1}$ randomly.
repeat
  // compute both gradients simultaneously
  $\eta_T \leftarrow \eta_T - \gamma g_T$ and $\eta_C \leftarrow \eta_C - \gamma g_C$
until convergence
Output: η_T, η_C

Algorithm 1 Following Gradients in Inverse-Weighted Games (scalar form)

Input: Choice of value functions V_F, V_G, normalization function h, learning rate γ.
Initialize $\eta_{T,1:K-1}$ and $\eta_{C,1:K-1}$ randomly. $g_T[t] = 0, g_C[t] = 0$ for $t = 1, \ldots, K-1$.
repeat
  // for each parameter
  for $t = 1$ **to** $K-1$ **do**
    // for each game
    for $k = 1$ **to** $K-1$ **do**
      // grad. from each game, scalar form of Equation (7)
      $g_T[t] = g_T[t] + \frac{dh(\eta)_k}{d\eta_t}\Big|_{\eta_T} \cdot \frac{dV_F^k}{d\theta_k}\Big|_{h(\eta_T),h(\eta_C)}$
      $g_C[t] = g_C[t] + \frac{dh(\eta)_k}{d\eta_t}\Big|_{\eta_C} \cdot \frac{dV_G^k}{d\theta_k}\Big|_{h(\eta_T),h(\eta_C)}$
    end for
  end for
  // Move toward solution where all games have zero gradient
  $\eta_T \leftarrow \eta_T - \gamma g_T$ and $\eta_C \leftarrow \eta_C - \gamma g_C$
  $g_T[t] = 0$ and $g_C[t] = 0$ for $t = 1, \ldots, K-1$
until convergence
Output: η_T, η_C

G Experiments

G.1 Data

Gamma Simulation We draw x from a 32D multivariate normal $\mathcal{N}(0, 10I)$. We simulate conditionally gamma failure times with mean μ_t a log-linear function of x with coefficients for each feature drawn Unif$(0, 0.1)$. The censoring times are also conditionally gamma with mean $0.9 * \mu_t$. Both distributions have constant variance 0.05. α, β parameterization of the gamma is recovered from mean, variance by $\alpha = \mu^2/\sigma^2$ and $\beta = \mu/\sigma^2$. T and C are conditionally independent given X. Each random seed draws a new dataset.

We report metrics as a function of training size. We use training sizes [200,400,600,800,1000]. We use validation size 1024 and testing size 2048.

Survival MNIST Survival-MNIST [Gensheimer, 2019, Pölsterl, 2019] draws times conditionally on MNIST label Y. This means digits define risk groups and $T \perp\!\!\!\perp X \mid Y$. Times within a digit are i.i.d. The model only sees the image pixels X as covariates so it must learn to classify digits (risk groups) to model times. The PyCox package [Kvamme et al., 2019] uses Exponential times. We follow Goldstein et al. [2020] and use Gamma times. T's mean is $10 * (Y + 1)$ so that lower labels Y mean sooner event times. We set the variance constant to 0.05. C is drawn similarly but with $9.9 * (Y + 1)$. Each random seed draws a new dataset.

We report metrics as a function of training size. We use training sizes [512, 1024, 2048, 4096, 8192, 10240]. We use validation size 1024 and testing size 2048.

Real Data We report results on

- SUPPORT [Knaus et al., 1995] which includes severely ill hospital patients. There are 14 features. we split into 5,323 for training, 1774 for validation, and 1776 for testing.

- METABRIC [Curtis et al., 2012]. There are 9 features. We split into 1,142 for training, 380 for validation, and 382 for testing.
- ROTT [Foekens et al., 2000] and GBSG [Schumacher et al., 1994] combined into one dataset (ROTT. & GBSG). There are 7 features. We split into 1,339 for training, 446 for validation, and 447 for testing.

For more description see Therneau [2021], Katzman et al. [2018], Chen [2020].

In the main text, we report results on a subset of these datasets with metrics as a function of training size. We use training sizes [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 175, 200]. We use validation size 300 and always use the entire testing set. We standardize all real data with the training set mean and standard deviation.

G.2 Models

In all experiments except for MNIST, we use a 3-hidden-layer ReLU network. The hidden sizes are [128, 64, 64] for the Gamma simulation and [128,256,64] for the real data. We output 20 categorical bins. See Appendix H.1 for different choices of number of bins, which did not show any significant differences in results. For MNIST we first use a small convolutional network and follow with the same fully-connected network, but using hidden sizes [512,256,64].

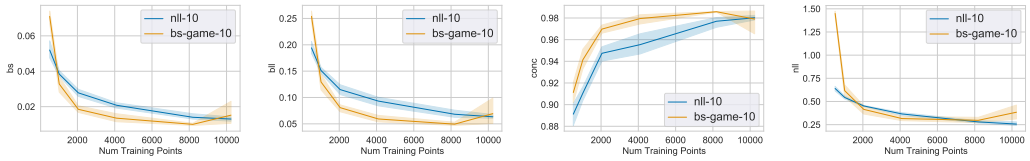
G.3 Training

We use learning rate 0.001 in all experiments for all losses using the Adam optimizer. We train for 300 epochs for the simulated data and 200 for the real data. For all data and all losses, this was enough to overfit on the training data. We use no weight decay or dropout.

H Ablations

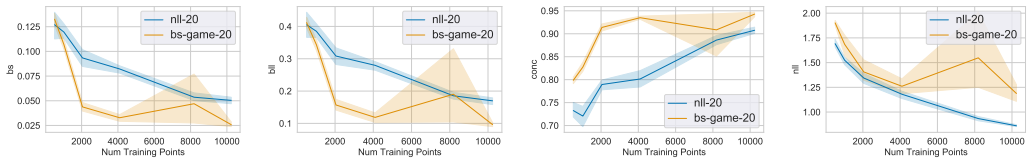
H.1 Changing number of bins on MNIST

Changing number of categorical bins (K) in [10,20,30,40,50]. Cannot directly compare between two choices of K due to changing meaning of likelihood/BS/Concordance but can compare NLL and BS-Game at each K. Trends similar across all choices of K.



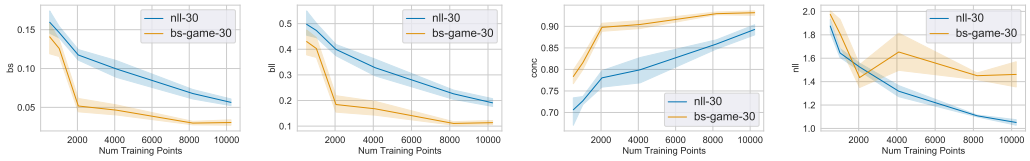
(a) uncens. BS (b) uncens. BLL (c) concordance (d) cat NLL

Figure 9: 10 bins. NLL (Blue). BS-Game (Orange).



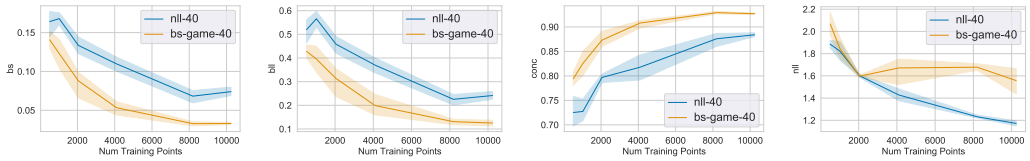
(a) uncens. BS (b) uncens. BLL (c) concordance (d) cat NLL

Figure 10: 20 bins. NLL (Blue). BS-Game (Orange).



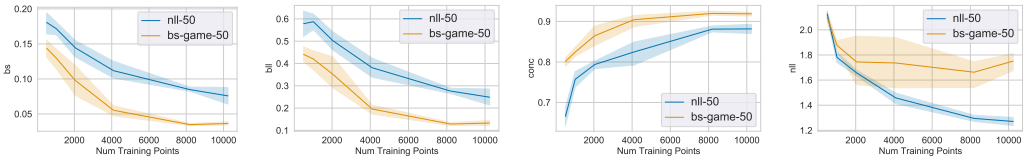
(a) uncens. BS (b) uncens. BLL (c) concordance (d) cat NLL

Figure 11: 30 bins. NLL (Blue). BS-Game (Orange).



(a) uncens. BS (b) uncens. BLL (c) concordance (d) cat NLL

Figure 12: 40 bins. NLL (Blue). BS-Game (Orange).



(a) uncens. BS (b) uncens. BLL (c) concordance (d) cat NLL

Figure 13: 50 bins. NLL (Blue). BS-Game (Orange).

I IPCW games at have a stationary point at data distributions

Setup. We define an *Inverse-Weighted Survival Game* at time t by IPCW estimators $L^{cw,t}$ of time-dependent loss L^t, where the failure and censor models re-weight each other's loss:

$$V_F^t(\theta) = L^{cw,t}_{F^\star,G^\star}(F_{\theta_T}; G_{\theta_C}), \quad V_G^t(\theta) = L^{cw,t}_{G^\star,F^\star}(G_{\theta_C}; F_{\theta_T})$$

In the following, it will be useful to separate the parameters at time t that are being optimized. We use $\overline{\theta_{Tt}}$ to mean all of the parameters in θ_T except for θ_{Tt} i.e. $\theta_{T1}, \ldots, \theta_{T,t-1}, \theta_{T,t+1}, \ldots, \theta_{T,K-1}$ and use analogous notation for θ_C. In place of $V_F^t(\theta)$ and $V_G^t(\theta)$ we write $V_F^t(\theta_{Tt}; \overline{\theta_{Tt}}, \theta_C)$ and $V_G^t(\theta_{Ct}; \overline{\theta_{Ct}}, \theta_T)$ where everything to the right of the semi-colon is not differentiated with respect to.

Proposition. Assume $\exists \theta_T^\star \in \Theta_T, \exists \theta_C^\star \in \Theta_C$ such that $F^\star = F_{\theta_T^\star}$ and $G^\star = G_{\theta_C^\star}$. Assume $L^t_{P^\star}(P)$ is proper, i.e. for model P and sampling distribution $P^\star$, L^t satisfies $P \neq P^\star \implies L^t_{P^\star}(P^\star) \leq L^t_{P^\star}(P)$. Then $(\theta_T^\star, \theta_C^\star)$ is a stationary point of the game Equation (4) for any t.

Proof. Choose an arbitrary t. We need to plug $(\theta_T^\star, \theta_C^\star)$ into the value functions V_F^t, V_G^t and show that

$$(\nabla_{\theta_{Tt}} V_F^t(\theta_{Tt}; \overline{\theta_{Tt}}, \theta_C))\big|_{\theta_T^\star, \theta_C^\star} = 0$$
$$(\nabla_{\theta_{Ct}} V_G^t(\theta_{Ct}; \overline{\theta_{Ct}}, \theta_T))\big|_{\theta_T^\star, \theta_C^\star} = 0$$

First, by definition,

$$V_F^t(\theta^\star_{Tt}; \overline{\theta^\star_{Tt}}, \theta^\star_C) = L^{cw,t}_{F^\star,G^\star}(F_{\theta_T^\star}; G_{\theta_C^\star})$$
$$V_G^t(\theta^\star_{Ct}; \overline{\theta^\star_{Ct}}, \theta^\star_T) = L^{cw,t}_{G^\star,F^\star}(G_{\theta_C^\star}; F_{\theta_T^\star})$$

Then by $F^\star = F_{\theta_T^\star}$ and $G^\star = G_{\theta_C^\star}$ this means

$$V_F^t(\theta^\star_{Tt}; \overline{\theta^\star_{Tt}}, \theta^\star_C) = L^{cw,t}_{F^\star,G^\star}(F_{\theta_T^\star}; G^\star)$$
$$V_G^t(\theta^\star_{Ct}; \overline{\theta^\star_{Ct}}, \theta^\star_T) = L^{cw,t}_{G^\star,F^\star}(G_{\theta_C^\star}; F^\star)$$

Next by the IPCW identity when the re-weighting distribution is correct this means

$$V_F^t(\theta^\star_{Tt}; \overline{\theta^\star_{Tt}}, \theta^\star_C) = L^t_{F^\star}(F_{\theta_T^\star})$$
$$V_G^t(\theta^\star_{Ct}; \overline{\theta^\star_{Ct}}, \theta^\star_T) = L^t_{G^\star}(G_{\theta_C^\star})$$

$V_F^t(\cdot;\cdot,\theta_C^\star)$ and $L^t_{F^\star}(F_{(\cdot)})$ are everywhere the same function of θ_T. Likewise $V_G^t(\cdot;\cdot,\theta_T^\star)$ and $L^t_{G^\star}(G_{(\cdot)})$ are everywhere the same function of θ_C. Therefore their gradients are equal everywhere:

$$\nabla_{\theta_{Tt}} V_F^t(\cdot;\cdot,\theta_C^\star) = \nabla_{\theta_{Tt}} L^t_{F^\star}$$
$$\nabla_{\theta_{Ct}} V_G^t(\cdot;\cdot,\theta_T^\star) = \nabla_{\theta_{Ct}} L^t_{G^\star}$$

So to verify the stationary point condition we only need to show that

$$(\nabla_{\theta_{Tt}} L^t_{F^\star})\big|_{\theta_T^\star} = 0$$
$$(\nabla_{\theta_{Ct}} L^t_{G^\star})\big|_{\theta_C^\star} = 0$$

Since L^t is proper we know that $L^t_{F^\star}(F_{\theta_T^\star}) = L^t_{F^\star}(F^\star) \leq L^t_{F^\star}(F_{\theta_T'})$ for any $\theta_T' \neq \theta^\star$ which implies the gradient condition for V_F^t. Analogous reasoning for V_G^t concludes the proof. □

J Proof That Collection of Discrete Brier Games have their only mutual stationary point at true solution

We show by induction on the time t of the IPCW BS game that the simultaneous gradient equations are only satisfied at $\hat{\theta_T} = \theta^\star_T$ and $\hat{\theta}_C = \theta^\star_C$. There is a lot of arithmetic but eventually it comes down to (1) substitution of one variable for another (2) assuming all previous timestep parameters are correct (induction) (3) finding the zeros of a quadratic (4) showing that one of the two solutions is the correct parameter and the other is invalid.

Note: this proof uses the notation that $\hat{\theta}$ is a model parameter and $\theta^\star$ is the correct one.

J.1 BS(1) (base case)

We can compute the expectations defining F-BS-CW(1) and G-BS-CW(1) in closed form. That gives us:

$$\text{F-BS-CW}(1) = \theta^\star_{T1}(1-\hat{\theta}_{T1})^2 + (1-\theta^\star_{T1})(1-\theta^\star_{C1})\frac{\hat{\theta}^2_{T1}}{1-\hat{\theta}_{C1}}$$
$$\text{G-BS-CW}(1) = \frac{\theta^\star_{C1}(1-\theta^\star_{T1})(1-\hat{\theta}_{C1})^2}{1-\hat{\theta}_{T1}} + (1-\theta^\star_{T1})(1-\theta^\star_{C1})\frac{\hat{\theta}^2_{C1}}{1-\hat{\theta}_{T1}}$$

The derivatives are

$$\frac{d\text{F-BS-CW}(1)}{d\hat{\theta}_{T1}} = 2\frac{(1-\theta^\star_{T1})(1-\theta^\star_{C1})}{1-\hat{\theta}_{C1}}\hat{\theta}_{T1} - 2(1-\hat{\theta}_{T1})\theta^\star_{T1} = 0$$
$$\frac{d\text{G-BS-CW}(1)}{d\hat{\theta}_{C1}} = 2\frac{(1-\theta^\star_{T1})(1-\theta^\star_{C1})}{1-\hat{\theta}_{T1}}\hat{\theta}_{C1} - 2\frac{(1-\theta^\star_{T1})(1-\hat{\theta}_{C1})\theta^\star_{C1}}{1-\hat{\theta}_{T1}} = 0$$

We can take each derivative equation and write one variable in terms of the other. First, taking $d\text{F-BS-CW}/d\hat{\theta}_{T1}$ and writing $\hat{\theta}_{T1}$ in terms of $\hat{\theta}_{C1}$:

$$\frac{d\text{F-BS-CW}(1)}{d\hat{\theta}_{T1}} = 2\frac{(1-\theta^\star_{T1})(1-\theta^\star_{C1})}{1-\hat{\theta}_{C1}}\hat{\theta}_{T1} - 2(1-\hat{\theta}_{T1})\theta^\star_{T1} = 0$$

implies

$$\frac{(1-\theta^\star_{T1})(1-\theta^\star_{C1})}{1-\hat{\theta}_{C1}}\hat{\theta}_{T1} = (1-\hat{\theta}_{T1})\theta^\star_{T1}$$
$$\frac{(1-\theta^\star_{T1})(1-\theta^\star_{C1})}{1-\hat{\theta}_{C1}}\hat{\theta}_{T1} + \theta^\star_{T1}\hat{\theta}_{T1} = \theta^\star_{T1}$$
$$\Big(\frac{(1-\theta^\star_{T1})(1-\theta^\star_{C1})}{1-\hat{\theta}_{C1}} + \theta^\star_{T1}\Big)\hat{\theta}_{T1} = \theta^\star_{T1}$$
$$\hat{\theta}_{T1} = \frac{\theta^\star_{T1}}{\Big(\frac{(1-\theta^\star_{T1})(1-\theta^\star_{C1})}{1-\hat{\theta}_{C1}} + \theta^\star_{T1}\Big)}$$

The solution for $\hat{\theta}_{T1}$ is linear in $\hat{\theta}_{C1}$. Now solving for $\hat{\theta}_{C1}$ in the G-BS-CS derivative:

$$\frac{d\text{G-BS-CW}(1)}{d\hat{\theta}_{C1}} = 2\frac{(1-\theta^\star_{T1})(1-\theta^\star_{C1})}{1-\hat{\theta}_{T1}}\hat{\theta}_{C1} - 2\frac{(1-\theta^\star_{T1})(1-\hat{\theta}_{C1})\theta^\star_{C1}}{1-\hat{\theta}_{T1}} = 0$$

implies

$$\frac{(1-\theta^\star_{T1})(1-\theta^\star_{C1})}{1-\hat{\theta}_{T1}}\hat{\theta}_{C1} = \frac{(1-\theta^\star_{T1})(1-\hat{\theta}_{C1})}{1-\hat{\theta}_{T1}}\theta^\star_{C1}$$

Given $1-\theta^\star_{T1} \neq 0$ and $1-\hat{\theta}_{T1} \neq 0$, we have

$$(1-\theta^\star_{C1})\hat{\theta}_{C1} = (1-\hat{\theta}_{C1})\theta^\star_{C1}$$

which gives us $\hat{\theta}_{C1} = \theta^\star_{C1}$. Given $1 - \theta^\star_{T1} \neq 0$ and $1 - \hat{\theta}_{T1} \neq 0$, the above derivative equations jointly imply

$$\hat{\theta}_{T1} = \frac{\theta^\star_{T1}}{\left(\frac{(1-\theta^\star_{T1})(1-\theta^\star_{C1})}{1-\hat{\theta}_{C1}} + \theta^\star_{T1}\right)}, \quad \hat{\theta}_{C1} = \theta^\star_{C1}$$

Substituting $\hat{\theta}_{C1} = \theta^\star_{C1}$ in the formula for $\hat{\theta}_{T1}$ in terms of $\hat{\theta}_{C1}$, we have

$$\hat{\theta}_{T1} = \frac{\theta^\star_{T1}}{\left(\frac{(1-\theta^\star_{T1})(1-\theta^\star_{C1})}{1-\theta^\star_{C1}} + \theta^\star_{T1}\right)} = \frac{\theta^\star_{T1}}{(1-\theta^\star_{T1}) + \theta^\star_{T1}} = \theta^\star_{T1}$$

Therefore, under the assumptions, for the BS(1) case, we have the only stationary point at the two true 1st-timestep parameters: $\hat{\theta}_{T1} = \theta^\star_{T1}$ and $\hat{\theta}_{C1} = \theta^\star_{C1}$.

J.2 Induction step

We can proceed by induction over timesteps. Claim: given $P_\theta(T \leq a) = P^\star(T \leq a)$ and $P_\theta(C \leq a) = P^\star(C \leq a), \quad a = 1, \ldots, k$, the stationary point of the game BS(k+1) has to satisfy $P_\theta(T = k+1) = P^\star(T = k+1)$ and $P_\theta(C = k+1) = P^\star(C = k+1)$ i.e. $\hat{\theta}_{T,k+1} = \theta^\star_{T,k+1}$ and $\hat{\theta}_{C,k+1} = \theta^\star_{C,k+1}$.We first simplify F-BS-CW.

$$\text{F-BS-CW}(k+1) = \mathop{\mathbb{E}}_{T,C}\left[\frac{(1 - F_\theta(k+1))^2 \mathbb{1}\left[T \leq C\right] \mathbb{1}\left[U \leq k+1\right]}{P_\theta(C' \geq U)} + \frac{F_\theta(k+1)^2 \mathbb{1}\left[U > k+1\right]}{P_\theta(C' > k+1)}\right]$$

We simplify each term of F-BS-CW separately. The left term of F-BS-CW is

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{T,C} \frac{(1-F_\theta(k+1))^2 \mathbb{1}\left[T \le C\right] \mathbb{1}\left[U \le k+1\right]}{P_\theta(C' \ge U)} \\
=&P_\theta(T > k+1)^2 \mathop{\mathbb{E}}_{T,C} \frac{\mathbb{1}\left[T \le C\right] \mathbb{1}\left[U \le k+1\right]}{P_\theta(C' \ge U)} \\
=&P_\theta(T > k+1)^2 \sum_{a=1}^{K}\sum_{b=1}^{K} P^\star(T=a)P^\star(C=b) \frac{\mathbb{1}\left[a \le b\right] \mathbb{1}\left[\min(a,b) \le k+1\right]}{P_\theta(C' \ge \min(a,b))} \\
&\qquad \Big[\text{condition } \mathbb{1}\left[a \le b\right] \text{moves from indicator to sum limits and } \min(a,b)=a\Big] \\
=&P_\theta(T > k+1)^2 \sum_{a=1}^{K}\sum_{b=a}^{K} \frac{P^\star(T=a)P^\star(C=b)\mathbb{1}\left[a \le k+1\right]}{P_\theta(C' \ge a)} \\
&\qquad \Big[\text{condition } \mathbb{1}\left[a \le k+1\right] \text{ moves from indicator to sum limit}\Big] \\
=&P_\theta(T > k+1)^2 \sum_{a=1}^{k+1}\sum_{b=a}^{K} \frac{P^\star(T=a)P^\star(C=b)}{P_\theta(C' \ge a)} \\
=&P_\theta(T > k+1)^2 \sum_{a=1}^{k+1} P^\star(T=a) \sum_{b=a}^{K} \frac{P^\star(C=b)}{P_\theta(C' \ge a)} \\
=&P_\theta(T > k+1)^2 \sum_{a=1}^{k+1} P^\star(T=a) \frac{P^\star(C \ge a)}{P_\theta(C' \ge a)} \\
&\left[\text{induction hypothesis: } P_\theta(C \le a) = P^\star(C \le a), \quad a=1,\dots,k \implies P_\theta(C > a) = P^\star(C > a), \quad a=1,\dots,k\right] \\
=&P_\theta(T > k+1)^2 \sum_{a=1}^{k+1} P^\star(T=a) \cdot 1 \\
=&P_\theta(T > k+1)^2 P^\star(T \le k+1) \\
=&(1-\sum_{i=1}^{k}\hat{\theta}_{Ti} - \hat{\theta}_{T(k+1)})^2 \sum_{i=1}^{k+1}\theta^\star_{Ti} \\
&\qquad \Big[\text{induction hypothesis: } P_\theta(T \le a) = P^\star(T \le a), \quad a=1,\dots,k\Big] \\
=&(1-\sum_{i=1}^{k}\theta^\star_{Ti} - \hat{\theta}_{T(k+1)})^2 \sum_{i=1}^{k+1}\theta^\star_{Ti} \\
=&(1-p-x)^2(p+t) \\
\triangleq& A, \qquad \text{where } p=\sum_{i=1}^{k}\theta^\star_{Ti}, q=\sum_{i=1}^{k}\theta^\star_{Ci}, x=\hat{\theta}_{T(k+1)}, y=\hat{\theta}_{C(k+1)}., t=\theta^\star_{T(k+1)} c=\theta^\star_{C(k+1)}.
\end{aligned}
$$

The right term of F-BS-CW is

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{T,C} \frac{F_\theta(k+1)^2 \mathbb{1}\left[U > k+1\right]}{P_\theta(C' > k+1)} \\
&= \frac{F_\theta(k+1)^2}{P_\theta(C' > k+1)} \mathop{\mathbb{E}}_{T,C} \mathbb{1}\left[U > k+1\right] \\
&\qquad \left[T \text{ and } C \text{ are independent means } \mathbb{1}\left[U > z\right] = \mathbb{1}\left[T > z\right] \mathbb{1}\left[C > z\right]\right] \\
&= \frac{F_\theta(k+1)^2}{P_\theta(C' > k+1)} P^\star(T > k+1) P^\star(C > k+1) \\
&= \frac{(\sum_{i=1}^{k+1} \hat{\theta}_{Ti})^2}{1 - \sum_{i=1}^{k+1} \hat{\theta}_{Ci}} (1 - \sum_{i=1}^{k+1} \theta^\star_{Ti})(1 - \sum_{i=1}^{k+1} \theta^\star_{Ci}) \\
&\quad \left[\text{induction hypothesis: } P_\theta(T \le a) = P^\star(T \le a) \text{ and } P_\theta(C \le a) = P^\star(C \le a), \quad a = 1, \ldots, k\right] \\
&= \frac{(\sum_{i=1}^{k} \theta^\star_{Ti} + \hat{\theta}_{T(k+1)})^2}{1 - \sum_{i=1}^{k} \theta^\star_{Ci} - \hat{\theta}_{C(k+1)}} (1 - \sum_{i=1}^{k+1} \theta^\star_{Ti})(1 - \sum_{i=1}^{k+1} \theta^\star_{Ci}) \\
&= \frac{(p+x)^2}{1-q-y}(1-p-t)(1-q-c) \triangleq B
\end{aligned}
$$

where again $p = \sum_{i=1}^{k} \theta^\star_{Ti}, q = \sum_{i=1}^{k} \theta^\star_{Ci}, x = \hat{\theta}_{T(k+1)}, y = \hat{\theta}_{C(k+1)}, t = \theta^\star_{T(k+1)}, c = \theta^\star_{C(k+1)}$.
To summarize, F-BS-CW$(k+1) = A + B$:

$$
\text{F-BS-CW}(k+1) = (1-p-x)^2(p+t) + \frac{(p+x)^2}{1-q-y}(1-p-t)(1-q-c)
$$

Then we simplify G-BS-CW.

$$
\text{G-BS-CW}(k+1) = \mathop{\mathbb{E}}_{T,C}\Big[\frac{(1-G_\theta(k+1))^2 \mathbb{1}\left[C < T\right] \mathbb{1}\left[U \le k+1\right]}{P_\theta(T' > U)} + \frac{G_\theta(k+1)^2 \mathbb{1}\left[U > k+1\right]}{P_\theta(T' > k+1)}\Big]
$$

The left term of G-BS-CW

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{T,C} \frac{(1-G_\theta(k+1))^2 \mathbb{1}\left[C<T\right] \mathbb{1}\left[U \le k+1\right]}{P_\theta(T'>U)} \\
=&(1-G_\theta(k+1))^2 \mathop{\mathbb{E}}_{T,C} \frac{\mathbb{1}\left[C<T\right] \mathbb{1}\left[U \le k+1\right]}{P_\theta(T'>U)} \\
=&(1-G_\theta(k+1))^2 \sum_{a=1}^{K}\sum_{b=1}^{K} P^\star(C=a)P^\star(T=b)\frac{\mathbb{1}\left[a<b\right] \mathbb{1}\left[\min(a,b) \le k+1\right]}{P_\theta(T'>\min(a,b))} \\
&\qquad \text{condition } \mathbb{1}\left[a<b\right] \text{ moves from indicator to sum limits and } \min(a,b)=a \\
=&(1-G_\theta(k+1))^2 \sum_{a=1}^{K}\sum_{b=a+1}^{K} \frac{P^\star(C=a)P^\star(T=b)\mathbb{1}\left[a \le k+1\right]}{P_\theta(T'>a)} \\
&\qquad \text{condition } \mathbb{1}\left[a \le k+1\right] \text{ moves from indicator to sum limits} \\
=&(1-G_\theta(k+1))^2 \sum_{a=1}^{k+1}\sum_{b=a+1}^{K} \frac{P^\star(C=a)P^\star(T=b)}{P_\theta(T'>a)} \\
&\qquad \left[\text{split sum over a into two terms: 1 through k, and k+1, recall b starts at a+1}\right] \\
=&(1-G_\theta(k+1))^2 \left(\sum_{a=1}^{k}\sum_{b=a+1}^{K} \frac{P^\star(C=a)P^\star(T=b)}{P_\theta(T'>a)} + \sum_{b=k+2}^{K} \frac{P^\star(C=k+1)P^\star(T=b)}{P_\theta(T'>k+1)}\right) \\
=&(1-G_\theta(k+1))^2 \left(\sum_{a=1}^{k} P^\star(C=a) \sum_{b=a+1}^{K} \frac{P^\star(T=b)}{P_\theta(T'>a)} + P^\star(C=k+1) \sum_{b=k+2}^{K} \frac{P^\star(T=b)}{P_\theta(T'>k+1)}\right) \\
=&(1-G_\theta(k+1))^2 \left(\sum_{a=1}^{k} \frac{P^\star(C=a)P^\star(T \ge a+1)}{P_\theta(T'>a)} + \frac{P^\star(C=k+1)P^\star(T>k+1)}{P_\theta(T'>k+1)}\right) \\
=&(1-G_\theta(k+1))^2 \left(\sum_{a=1}^{k} \frac{P^\star(C=a)P^\star(T>a)}{P_\theta(T'>a)} + \frac{P^\star(C=k+1)P^\star(T>k+1)}{P_\theta(T'>k+1)}\right) \\
&\left[\text{induction hypothesis: } P_\theta(T \le a) = P^\star(T \le a), \quad a=1,\dots,k \implies P_\theta(T>a) = P^\star(T>a), \quad a=1,\dots,k\right] \\
=&(1-G_\theta(k+1))^2 \left(\sum_{a=1}^{k} P^\star(C=a) + \frac{P^\star(C=k+1)P^\star(T>k+1)}{P_\theta(T'>k+1)}\right) \\
=&(1-\sum_{i=1}^{k}\hat{\theta}_{Ci} - \hat{\theta}_{C(k+1)})^2 \left(\sum_{i=1}^{k}\theta^\star_{Ci} + \frac{\theta^\star_{C(k+1)}(1-\theta^\star_{T(k+1)} - \sum_{i=1}^{k}\theta^\star_{Ti})}{1-\sum_{i=1}^{k}\hat{\theta}_{Ti} - \hat{\theta}_{T(k+1)}}\right) \\
&\left[\text{induction hypothesis: } P_\theta(T \le a) = P^\star(T \le a) \quad \text{and} \quad P_\theta(C \le a) = P^\star(C \le a), \quad a=1,\dots,k\right] \\
=&(1-\sum_{i=1}^{k}\theta^\star_{Ci} - \hat{\theta}_{C(k+1)})^2 \left(\sum_{i=1}^{k}\theta^\star_{Ci} + \frac{\theta^\star_{C(k+1)}(1-\theta^\star_{T(k+1)} - \sum_{i=1}^{k}\theta^\star_{Ti})}{1-\sum_{i=1}^{k}\theta^\star_{Ti} - \hat{\theta}_{T(k+1)}}\right) \\
=&(1-q-y)^2(q+\frac{c(1-t-p)}{1-p-x}) \triangleq C
\end{aligned}
$$

By symmetry with B, the right term is

$$
\mathop{\mathbb{E}}_{T,C} \frac{G_\theta(k+1)^2 \mathbb{1}\left[U>k+1\right]}{P_\theta(T'>k+1)} = \frac{(q+y)^2}{1-p-x}(1-q-c)(1-p-t) \triangleq D
$$

Again using $p = \sum_{i=1}^{k} \theta^\star_{Ti}, q = \sum_{i=1}^{k} \theta^\star_{Ci}, x = \hat{\theta}_{T(k+1)}, y = \hat{\theta}_{C(k+1)}, t = \theta^\star_{T(k+1)}, c = \theta^\star_{C(k+1)}$, we have

$$\begin{aligned}\text{G-BS-CW}(k+1) &= C + D \\ &= (1-q-y)^2(q + \frac{c(1-t-p)}{1-p-x}) + \frac{(q+y)^2}{1-p-x}(1-q-c)(1-p-t)\end{aligned}$$

The stationary point satisfies

$$\begin{aligned}\frac{\partial \text{G-wt-FBS(k+1)}}{\partial x} &= \frac{\partial A}{\partial x} + \frac{\partial B}{\partial x} \\ &= -2(1-p-x)(p+t) + 2\frac{(p+x)}{1-q-y}(1-p-t)(1-q-c) \\ &= 0 \\ \frac{\partial \text{F-wt-GBS(k+1)}}{\partial y} &= \frac{\partial C}{\partial y} + \frac{\partial D}{\partial y} \\ &= -2(1-q-y)(q + \frac{c(1-t-p)}{1-p-x}) + 2\frac{(q+y)}{1-p-x}(1-q-c)(1-p-t) = 0\end{aligned}$$

It's a system of quadratic equations with two unknowns. The system has analytical solutions. Solving the above equations for x, y by *Mathematica* (it is quite a long derivation manually), the solutions are

$$x = t, y = c$$

or

$$\begin{aligned}x =&(1/(-q+q^2+qc))(cp - qcp - qt + q^2t + ct \\ &- (p(-1+q+c+qp-q^2p-cp+qt-q^2t-ct))/((-1+q)(p+t)) \\ &+ (qp(-1+q+c+qp-q^2p-cp+qt-q^2t-ct))/((-1+q)(p+t)) \\ &- (t(-1+q+c+qp-q^2p-cp+qt-q^2t-ct))/((-1+q)(p+t)) \\ &+ (qt(-1+q+c+qp-q^2p-cp+qt-q^2t-ct))/((-1+q)(p+t))) \\ y =&(-1+q+c+qp-q^2p-cp+qt-q^2t-ct)/((-1+q)(p+t))\end{aligned}$$

To check if this second solution is valid, it would need to be the case that $q + y < 1$ because we only consider $k + 1 < K$. If we ask mathematica to simplify q+y that satisfies the above solution, then this holds:

$$q + y = \frac{-1 + q - c(-1+p+t)}{(-1+q)(p+t)}$$

The numerator and the denominator are both negative. If $k + 1 < K$ (we know BS at K is 0 and also we only have K-1 parameters), the numerator minus denominator =

$$\begin{aligned}-1 + q - c(-1+p+t) - (-1+q)(p+t) &= (-1+q)(1-p-t) - c(-1+p+t) \\ &= (-1+q+c)(1-p-t) \\ &< 0\end{aligned}$$

Therefore,

$$\sum_{i=1}^{k} \theta^\star_{Ci} + \hat{\theta}_{C(k+1)} = q + y > 1$$

This is invalid. So

$$x = t, y = c$$

is the only solution, i.e., $\hat{\theta}_{T(k+1)} = \theta^\star_{T(k+1)}, \hat{\theta}_{C(k+1)} = \theta^\star_{C(k+1)}$. By induction, we conclude that

$$\hat{\theta}_{Ti} = \theta^\star_{Ti}, \hat{\theta}_{Ci} = \theta^\star_{Ci}, i = 1, \ldots, K-1$$

By $\hat{\theta}_{TK} = 1 - \sum_{i=1}^{K-1} \hat{\theta}_{Ti}$ and $\hat{\theta}_{CK} = 1 - \sum_{i=1}^{K-1} \hat{\theta}_{Ci}$, we have

$$\hat{\theta}_{TK} = \theta^\star_{TK}, \hat{\theta}_{CK} = \theta^\star_{CK}$$

Therefore,

$$\hat{\theta}_{Ti} = \theta^\star_{Ti}, \hat{\theta}_{Ci} = \theta^\star_{Ci}, i = 1, \ldots, K$$

is the only stationary point for the game.