Bregman projection for calibration estimation

Jae Kwang Kim

June 22nd, 2026
IMSI Workshop

Coauthors

Yonghyun Kwon

Assistant Professor
Department of Mathematics
Korea Military Academy

Ph.D. Iowa State University, 2024

Yumou Qiu

Associate Professor
School of Mathematical Sciences
Peking University

Ph.D. Iowa State University, 2014

A motivating problem: estimating a mean with auxiliary information

We observe i.i.d. data $\{(X_i, Y_i)\}_{i=1}^n$ and want to estimate

$$\theta = E(Y).$$

Twist: the population mean of the auxiliary variable is *known*,

$$\mu_X = E(X) \quad \text{(known)}.$$

- The naive estimator $\bar{Y}$ ignores the known μ_X.
- If X and Y are correlated, we can do better.
- This is a canonical **over-identified** problem: one parameter, two moment conditions.

Question

How should we combine the sample with the known μ_X? And is there a single principle behind the many available estimators?

The estimating equations

Idea (calibration / reweighting). Replace the uniform weights $1/n$ by data-driven weights $\omega_i > 0$ that *exactly* satisfy both sample moment conditions:

$$\sum_{i=1}^{n} \omega_i (Y_i - \theta) = 0, \qquad \sum_{i=1}^{n} \omega_i (X_i - \mu_X) = 0.$$

- Two equations, n unknown weights $\Rightarrow$ infinitely many solutions.
- **Which weights?** Choose the ω *closest* to the uniform anchor $1/n$.
- "Closest" in what sense? $\leftarrow$ need to choose an objective function for constrained optimization.

Motivating calibration estimation

- Estimator class

$$\hat{\theta}_\omega = \frac{\sum_{i=1}^n \omega_i y_i}{\sum_{i=1}^n \omega_i}$$

- Working model

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + e_i, \tag{1}$$

where $e_i \overset{iid}{\sim} (0, \sigma_e^2)$ and e_i is independent of $\boldsymbol{x}_i$.

- Under (1), the mean-squared error decomposes as

$$E\left[(\hat{\theta}_\omega - \theta)^2\right] = \frac{1}{\left(\sum_{i=1}^n \omega_i\right)^2}\left[\left(\mathbf{B}(\boldsymbol{\omega})^\top \boldsymbol{\beta}\right)^2 + \sigma_e^2 \|\boldsymbol{\omega}\|^2\right], \qquad \mathbf{B}(\boldsymbol{\omega}) \equiv \sum_{i \in S} \omega_i \left(\boldsymbol{x}_i - \boldsymbol{\mu}_x\right), \tag{2}$$

into a squared-bias term and a variance term.

How to determine ω?

- Worst-case minimax approach (Hirshberg et al., 2021):

$$\hat{\boldsymbol{\omega}} \in \arg\min_{\omega} \max_{\beta} E\left[(\hat{\theta}_{\omega} - \theta)^2\right]; \tag{3}$$

- Note that

$$\max_{\beta} E\left[(\hat{\theta}_{\omega} - \theta)^2\right] < \infty \iff \mathbf{B}(\boldsymbol{\omega}) = \mathbf{0}.$$

- The solution (3) can be obtained by a constrained optimization problem: minimize

$$Q(\boldsymbol{\omega}) = \frac{\sum_{i=1}^{n} \omega_i^2}{\left(\sum_{i=1}^{n} \omega_i\right)^2} \tag{4}$$

subject to

$$\sum_{i=1}^{n} \omega_i \left(\boldsymbol{x}_i - \boldsymbol{\mu}_x\right) = 0. \tag{5}$$

- Equivalently, we can add the normalization constraint into the calibration problem: Minimize

$$Q(\boldsymbol{\omega}) = \sum_{i=1}^{n} {\omega_i}^2 \tag{6}$$

subject to calibration (5) and normalization

$$\sum_{i=1}^{n} \omega_i = 1. \tag{7}$$

- Under (7), the objective function in (6) is equivalent to

$$D(\boldsymbol{\omega}) = \sum_{i=1}^{n} \left(\omega_i - \frac{1}{n} \right)^2 \tag{8}$$

which is the Euclidean distance under L_2 norm.

- The solution can be expressed as

$$\hat{\omega}_i = \frac{1}{n} + \hat{\lambda}_0 + \hat{\boldsymbol{\lambda}}_1^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_x)$$

  where $\hat{\lambda}_0$ and $\hat{\boldsymbol{\lambda}}_1$ solve the joint calibration equation (5) and (7).

- The final estimator is algebraically equivalent to the regression estimator (Anderson, 1957):

$$\sum_{i=1}^{n} \hat{\omega}_i y_i = \bar{y}_n + (\boldsymbol{\mu}_x - \bar{\boldsymbol{x}}_n)^\top \hat{\boldsymbol{\beta}}_1 = \hat{\beta}_0 + \boldsymbol{\mu}_x^\top \hat{\boldsymbol{\beta}}_1.$$

  This equivalence was first established by Fuller (1975) in the context of survey sampling.

Remark

- Two calibration methods can be developed in terms of dealing with normalization.
- **Method 1**: Apply the normalization later by using

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \hat{\omega}_i y_i}{\sum_{i=1}^{n} \hat{\omega}_i}$$

  where $\hat{\omega}_i$ is the unnormalized calibration minimizing $D(\boldsymbol{\omega}) = \sum_{i=1}^{n} \left(\omega_i - n^{-1}\right)^2$ subject to (5) only.
- **Method 2**: In addition to (5), include the normalization (7) into the calibration constraints.
- The two methods are not exactly the same, but are asymptotically equivalent to the first order.

- Method 1 can also be implemented using

$$\hat{\theta} \in \arg\min_{\theta} \min_{\omega \in \mathcal{C}(\theta)} D(\boldsymbol{\omega}) = \arg\min_{\theta} D\left(\hat{\boldsymbol{\omega}}(\theta)\right), \tag{9}$$

where

$$\mathcal{C}(\theta) = \left\{ \boldsymbol{\omega} \in \mathbb{R}^n; \sum_{i=1}^{n} \omega_i U(\theta; \boldsymbol{x}_i, y_i) = \mathbf{0} \right\}$$

and

$$U(\theta; \boldsymbol{x}, y) = \begin{pmatrix} y - \theta \\ \boldsymbol{x} - \boldsymbol{\mu}_x \end{pmatrix}.$$

- In (9), the second equality holds with

$$\hat{\boldsymbol{\omega}}(\theta) = \arg\min_{\omega \in \mathcal{C}(\theta)} D(\boldsymbol{\omega}).$$

Research Goals

In the class of calibration problem

1. Wish to address the selection bias in the sample
2. Wish to generalize the distance measure
3. Wish to achieve some optimality

Generalizing the distance measure for calibration

- Consider the finite population setup. Sample S is selected from the finite population.
- Let $\omega_i^{(0)}$ be the initial weight for unit i. (To be determined later)
- Let $G(\cdot)$ be a strictly convex and differentiable function.
- Deville and Särndal (1992) calibration: Minimize

$$D_{\mathrm{DS}}(\boldsymbol{\omega} \parallel \boldsymbol{\omega}^{(0)}) = \sum_{i \in S} \omega_i^{(0)} G\left(\omega_i / \omega_i^{(0)}\right) \tag{10}$$

subject to

$$\sum_{i \in S} \omega_i \tilde{\boldsymbol{x}}_i = \mathbf{0}, \tag{11}$$

where $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}}_N$ and $\bar{\boldsymbol{x}}_N$ is the finite population mean.

- To make $D_{\mathrm{DS}}(\boldsymbol{\omega} \parallel \boldsymbol{\omega}^{(0)}) \geq D_{\mathrm{DS}}(\boldsymbol{\omega}^{(0)} \parallel \boldsymbol{\omega}^{(0)})$, we require $G'(1) = 0$.

Proposal: Bregman entropy calibration (Kim et al., 2026)

- Find $\hat{\omega}_i$ by minimizing the Bregman divergence measure

$$D_G(\boldsymbol{\omega} \parallel \boldsymbol{\omega}^{(0)}) = \sum_{i \in S} \left\{ G(\omega_i) - G\left(\omega_i^{(0)}\right) - G'\left(\omega_i^{(0)}\right)\left(\omega_i - \omega_i^{(0)}\right) \right\} \tag{12}$$

  subject to the same calibration constraint in (11).

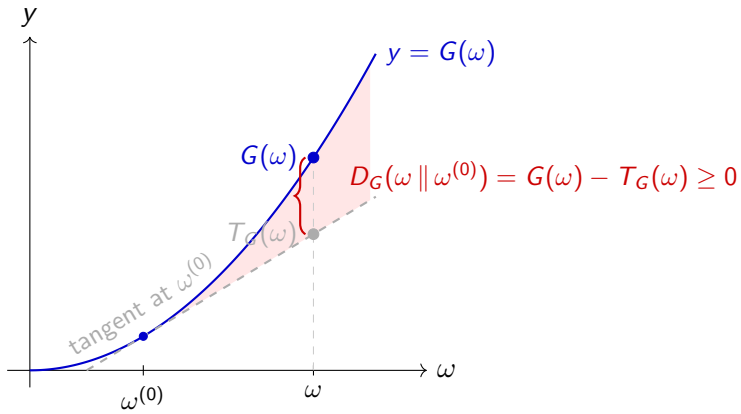
- The Bregman divergence $D_G(\omega_i \parallel \omega_i^{(0)})$ can be understood as the difference between $G(\omega_i)$ and its tangent line

$$T_G(\omega_i) = G(\omega_i^{(0)}) + G'(\omega_i^{(0)})(\omega_i - \omega_i^{(0)})$$

  at $\omega_i = \omega_i^{(0)}$.

- The strict convexity of G implies $D_G(\omega_i \parallel \omega_i^{(0)}) \geq 0$ with equality if and only if $\omega_i = \omega_i^{(0)}$.

Bregman Divergence: Geometric Interpretation



Computation

- By the Lagrangian multiplier method, we maximize

$$\mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\lambda}) = -\sum_{i \in S} \left\{ G(\omega_i) - G\left(\omega_i^{(0)}\right) - G'\left(\omega_i^{(0)}\right)\left(\omega_i - \omega_i^{(0)}\right) \right\} + \boldsymbol{\lambda}^\top \sum_{i \in S} \omega_i \tilde{\boldsymbol{x}}_i.$$

- Setting $\partial \mathcal{L} / \partial \omega_i = 0$ gives the KKT stationarity condition.
- Solving

$$\frac{\partial}{\partial \omega_i} \mathcal{L} = -G'(\omega_i) + G'\left(\omega_i^{(0)}\right) + \boldsymbol{\lambda}^\top \tilde{\boldsymbol{x}}_i = 0$$

for ω_i gives the form of *natural parametrization*:

$$G'(\omega_i) = G'(\omega_i^{(0)}) + \boldsymbol{\lambda}^\top \tilde{\boldsymbol{x}}_i. \tag{13}$$

- Thus, the solution can be expressed as

$$\omega_i^\star(\boldsymbol{\lambda}) = \left(G'\right)^{-1}\left\{G'(\omega_i^{(0)}) + \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\lambda}\right\} \tag{14}$$

The final weight is completely determined by $\boldsymbol{\lambda}$.

- Thus, we can express

$$\begin{aligned}
& \ell(\boldsymbol{\lambda}) \\
\equiv\ & \mathcal{L}\left\{\boldsymbol{\omega}^\star(\boldsymbol{\lambda}), \boldsymbol{\lambda}\right\} \\
=\ & -\sum_{i\in S}\left[G\left\{\omega_i^\star(\boldsymbol{\lambda})\right\} - G\left(\omega_i^{(0)}\right) - G'\left(\omega_i^{(0)}\right)\left(\omega_i^\star(\boldsymbol{\lambda}) - \omega_i^{(0)}\right)\right] + \boldsymbol{\lambda}^\top \sum_{i\in S}\omega_i^\star(\boldsymbol{\lambda})\,\tilde{\boldsymbol{x}}_i \\
=\ & \sum_{i\in S}\underbrace{\left\{G'(\omega_i^{(0)}) + \boldsymbol{\lambda}^\top\tilde{\boldsymbol{x}}_i\right\}\omega_i^\star(\boldsymbol{\lambda}) - G\left\{\omega_i^\star(\boldsymbol{\lambda})\right\}}_{=F\left(G'(\omega_i^{(0)}) + \tilde{\boldsymbol{x}}_i^\top\boldsymbol{\lambda}\right)} + \underbrace{\sum_{i\in S}\left\{G(\omega_i^{(0)}) - G'(\omega_i^{(0)})\omega_i^{(0)}\right\}}_{=\text{const.}}
\end{aligned}$$

to obtain $\hat{\boldsymbol{\lambda}} = \arg\min \ell(\boldsymbol{\lambda})$.

Dual problem formulation

- Convex conjugate function (Legendre transform) of G:

$$F(\nu) := \sup_{\omega \in V} \{\omega\nu - G(\omega)\} = \nu \cdot \omega^\star(\nu) - G\big(\omega^\star(\nu)\big),$$

where $\omega^\star(\nu) = (G')^{-1}(\nu)$.

- The convex conjugate function satisfies

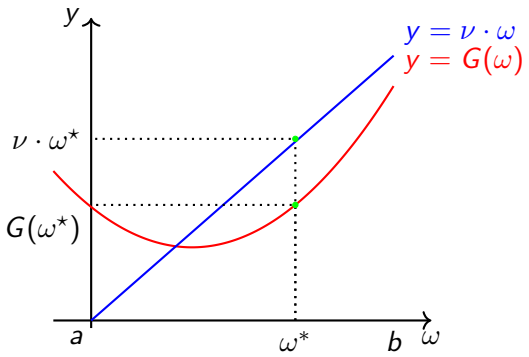
$$F'(\nu) = \big(G'\big)^{-1}(\nu) \quad \text{and} \quad G'(\omega) = \big(F'\big)^{-1}(\omega).$$

- Using the convex conjugate function, we can express

$$\ell(\boldsymbol{\lambda}) = \sum_{i \in S} F\big(G'(\omega_i^{(0)}) + \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\lambda}\big) + C,$$

where $C = \sum_{i \in S} \{G(\omega_i^{(0)}) - G'(\omega_i^{(0)})\omega_i^{(0)}\}$.

Legendre Transformation (Convex conjugate function)



The function $G(\omega)$ is defined on the interval $[a, b]$. The difference $\nu \cdot \omega - G(\omega)$ takes a maximum at $\omega^\star = \omega^\star(\nu)$. Thus, $\rho(\nu) = \nu \cdot \omega^\star - G(\omega^\star)$ is the Legendre transformation of G.

Examples

Generalized Entropy	$G(\omega)$	$F(\nu)$
Squared loss	$\omega^2/2$	$\nu^2/2$
Kullback-Leibler	$\omega\{\log(\omega)-1\}$	$\exp(\nu)$
Shifted KL	$(\omega-1)\{\log(\omega-1)-1\}$	$\nu+\exp(\nu)$
Empirical likelihood	$-\log(\omega)$	$-1-\log(-\nu)$
Squared Hellinger	$(\sqrt{\omega}-1)^2$	$\nu/(1-\nu)$
Rényi entropy $(\alpha \neq 0,-1)$	$\frac{1}{\alpha+1}\omega^{\alpha+1}$	$\frac{\alpha}{\alpha+1}\nu^{\frac{\alpha+1}{\alpha}}$

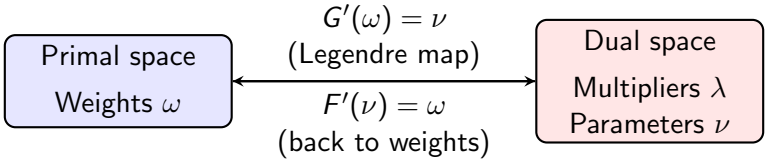
Table: Examples of generalized entropies, $G(\omega)$ and the corresponding convex conjugate function F

Duality

- Let
$$\nu_i(\boldsymbol{\lambda}) = G'(\omega_i^{(0)}) + \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\lambda}$$
be the parameter that characterizes the final weight
$$\omega_i^\star(\boldsymbol{\lambda}) = F'\left(\nu_i(\boldsymbol{\lambda})\right).$$

- Note that
$$G'\left\{\omega_i^\star(\boldsymbol{\lambda})\right\} = G'\left\{F'\left(\nu_i(\boldsymbol{\lambda})\right)\right\} = \nu_i(\boldsymbol{\lambda})$$
as $(G')^{-1} = F'$.

Primal–Dual Picture of Calibration



Intuition

- The primal problem minimizes divergence in the weight space.
- The dual problem minimizes divergence in the multiplier space.
- The Legendre transform provides the bridge: $\nu = G'(\omega)$.
- At the optimum, the two problems are equivalent and consistent.

Remark: Structural Parallel with GLM

- The calibrated weight map
$$\hat{\omega}_i = F'\big(G'(\omega_i^{(0)}) + \tilde{\boldsymbol{x}}_i^\top \hat{\boldsymbol{\lambda}}\big)$$
is structurally identical to a **generalized linear model (GLM)** with canonical link:

	GLM	Bregman Calibration
Mean parameter	$\mu_i = b'(\nu_i)$	$\hat{\omega}_i = F'(\nu_i)$
Natural parameter	$\nu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$	$\nu_i = G'(\omega_i^{(0)}) + \boldsymbol{x}_i^\top \boldsymbol{\lambda}$
Cumulant function	$b(\cdot)$	$F(\cdot)$
Canonical link	$(b')^{-1}(\cdot)$	$G'(\cdot)$
Offset	—	$G'(\omega_i^{(0)})$

- The convex conjugate F for weight calibration plays the role of the cumulant function b in the GLM, and $\omega_i^{(0)}$ enters as a unit-specific **offset** in the natural parameter.
- Just as the **canonical link** in a GLM is $(b')^{-1}$, the **calibration link** $g = G'$ maps weights to natural parameters, with its inverse $g^{-1} = F'$ mapping back.

- Let $\boldsymbol{\omega}^{(1)} \in \mathbb{R}^n$ be any element in

$$\mathcal{C} = \left\{ \boldsymbol{\omega} \in \mathbb{R}^n; \sum_{i \in S} \omega_i \tilde{\boldsymbol{x}}_i = \mathbf{0} \right\}$$

- Calibration equation can be viewed as a score equation for GLM:

$$\sum_{i=1}^{n} \left[\omega_i^{(1)} - F' \left\{ G'(\omega_i^{(0)}) + \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\lambda} \right\} \right] \tilde{\boldsymbol{x}}_i = \mathbf{0}.$$

  Thus, Bregman projection for calibration is equivalent to fitting a GLM of $\boldsymbol{\omega}^{(1)} \in \mathcal{C}$ onto the linear space of span $\left\{ \tilde{X}_1, \ldots, \tilde{X}_p \right\}$ using $G'(\cdot)$ as the canonical link function.

- See Cho et al. (2026) for more details on Bregman projection in statistics.

Pythagorean theorem

Theorem 1

Let

$$\mathcal{C} = \left\{ \boldsymbol{\omega} \in \mathbb{R}^n; \sum_{i \in S} \omega_i \tilde{\boldsymbol{x}}_i = \mathbf{0} \right\} \tag{15}$$

For any element $\boldsymbol{\omega} \in \mathcal{C}$*, we obtain*

$$D_G\left(\boldsymbol{\omega} \parallel \boldsymbol{\omega}^{(0)}\right) = D_G\left(\boldsymbol{\omega} \parallel \hat{\boldsymbol{\omega}}\right) + D_G\left(\hat{\boldsymbol{\omega}} \parallel \boldsymbol{\omega}^{(0)}\right), \tag{16}$$

where $\hat{\boldsymbol{\omega}}$ *is the vector of* $\omega_i^\star(\hat{\boldsymbol{\lambda}})$ *with* $\hat{\boldsymbol{\lambda}} = \arg\min_\lambda \ell(\boldsymbol{\lambda})$ *and* $\omega_i^\star(\boldsymbol{\lambda}) = F'\left(\nu_i(\boldsymbol{\lambda})\right)$.

Interpretation

- **Primal view (weight space).** The calibrated weights $\hat{\boldsymbol{\omega}}$ are the Bregman projection of $\boldsymbol{\omega}^{(0)}$ onto $\mathcal{C}$ in (15):

$$\hat{\boldsymbol{\omega}} \in \arg\min_{\omega \in \mathcal{C}} \sum_{i \in S} D_G(\omega_i \,\|\, \omega_i^{(0)}). \tag{17}$$

  Thus, the solution $\hat{\boldsymbol{\omega}}$ is the closest point to $\boldsymbol{\omega}^{(0)}$ in $D_G(\cdot \,\|\, \omega^{(0)})$ among all points in $\boldsymbol{\omega} \in \mathcal{C}$.

- **Dual view (multiplier space).** The Lagrange multiplier solves a low-dimensional convex problem:

$$\hat{\boldsymbol{\lambda}} \in \arg\min_{\lambda \in \mathbb{R}^p} \ \ell(\boldsymbol{\lambda}), \qquad \ell(\boldsymbol{\lambda}) = \sum_{i \in S} F\big(\nu_i(\boldsymbol{\lambda})\big) + C,$$

  where $\nu_i(\boldsymbol{\lambda}) = G'(\omega_i^{(0)}) + \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\lambda}$. In other words, $G'(\hat{\omega}_i) - G'(\omega_i^{(0)})$ lies in the span of the calibration variables $\tilde{\boldsymbol{x}}_i$.

- **Bridge (Legendre map).**

$$\hat{\omega}_i = F'\big(\nu_i(\hat{\boldsymbol{\lambda}})\big) \quad \text{and} \quad \nu_i(\hat{\boldsymbol{\lambda}}) = G'\left(\hat{\omega}_i\right). \tag{18}$$

A graphical illustration for $G(\omega) = \omega^2$: $\boldsymbol{\omega} \in \mathbb{R}^n$

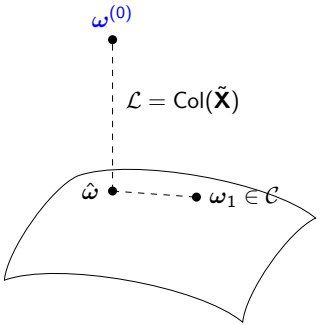


Figure: The divergence $D_G(\boldsymbol{\omega} \parallel \boldsymbol{\omega}^{(0)})$ among $\boldsymbol{\omega} \in \mathcal{C}_\omega$ is minimized at $\boldsymbol{\omega} = \hat{\boldsymbol{\omega}}$ where $\hat{\boldsymbol{\omega}}$ satisfies $\hat{\boldsymbol{\omega}} \in \mathcal{C}$ and $G'(\boldsymbol{\omega}^{(0)}) - G'(\hat{\boldsymbol{\omega}}) \in \mathcal{L} = \text{Col}\left(\tilde{\mathbf{X}}\right)$. The subspace C_ω is the class of the weights satisfying the calibration constraints. Note that $\boldsymbol{\omega}_1 - \hat{\boldsymbol{\omega}} \in \text{Ker}(\tilde{\mathbf{X}}^\top)$ is orthogonal to $G'(\boldsymbol{\omega}^{(0)}) - G'(\hat{\boldsymbol{\omega}}) \in \mathcal{L} = \text{Col}\left(\tilde{\mathbf{X}}\right)$.

Two calibration methods

- Deville and Särndal (1992) method: Minimize

$$\sum_{i \in S} \omega_i^{(0)} G\left(\omega_i / \omega_i^{(0)}\right) \tag{19}$$

subject to the calibration condition in (11).

- Our proposed method: Minimize

$$\sum_{i \in S} D_G(\omega_i \parallel \omega_i^{(0)})$$

subject to (11), where $D_G(\omega_i \parallel \omega_i^{(0)})$ is the Bregman divergence measure generated by G.

Remark

- Deville and Särndal (1992) method leads to

$$\hat{\omega}_i = \omega_i^{(0)} \times F'(\boldsymbol{x}_i^\top \hat{\boldsymbol{\lambda}}),$$

  where F is the convex conjugate function of G.
- Our proposed method leads to

$$\hat{\omega}_i = F' \left\{ G'(\omega_i^{(0)}) + \boldsymbol{x}_i^\top \hat{\boldsymbol{\lambda}} \right\}$$

  where $F'(\cdot) = \left(G'\right)^{-1}(\cdot)$.
- In the special case of exponential tilting (ET) weight (using KL divergence), we have $F(\nu) = \exp(\nu)$ and the two methods coincide.

Asymptotic results (under the design-based approach)

- Under some regularity conditions, the Deville-Särndal calibration estimator is asymptotically equivalent to the debiased regression estimator:

$$\hat{\theta}_{\text{dreg}} = \sum_{i \in S} \omega_i^{(0)} y_i - \sum_{i \in S} \omega_i^{(0)} \tilde{\boldsymbol{x}}_i^\top \hat{\boldsymbol{\beta}},$$

  where

$$\hat{\boldsymbol{\beta}} = \left\{ \sum_{i \in S} \omega_i^{(0)} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top \right\}^{-1} \sum_{i \in S} \omega_i^{(0)} \tilde{\boldsymbol{x}}_i y_i.$$

- For the case of $G(\omega) = \frac{1}{2}(\omega - 1)^2$, the equivalence is exact.
- The asymptotic equivalence holds regardless of the choice of G in the objective function (19) for calibration.

- On the other hand, our proposed calibration estimator is asymptotically equivalent to the following regression estimator

$$\hat{\theta}_{\text{greg}} = \sum_{i \in S} \omega_i^{(0)} y_i - \sum_{i \in S} \omega_i^{(0)} \tilde{\boldsymbol{x}}_i^\top \hat{\boldsymbol{\beta}}_G, \tag{20}$$

where

$$\hat{\boldsymbol{\beta}}_G = \left\{ \sum_{i \in S} \frac{\tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top}{G''(\omega_i^{(0)})} \right\}^{-1} \sum_{i \in S} \frac{\tilde{\boldsymbol{x}}_i y_i}{G''(\omega_i^{(0)})} \tag{21}$$

and $G''(\omega) = d^2 G(\omega)/d\omega^2$.

- In DS, the choice of G is irrelevant to efficiency. In Bregman calibration, G controls the regression weighting and hence the asymptotic variance.

Design-optimal estimation

- Under Poisson sampling, the design-optimal regression estimator (Montanari, 1987) is obtained with

$$\hat{\boldsymbol{\beta}}_{\text{opt}} = \left(\sum_{i \in S} q_i \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top \right)^{-1} \left(\sum_{i \in S} q_i \tilde{\boldsymbol{x}}_i y_i \right)$$

  where $q_i = \pi_i^{-2} - \pi_i^{-1}$.

- Since $\omega_i^{(0)} = \pi_i^{-1}$, by (21), we only need to find a special entropy function $G(\cdot)$ such that

$$1/G''(\pi_i^{-1}) = \pi_i^{-2} - \pi_i^{-1},$$

  which is satisfied with

$$G(\omega) = (\omega - 1)\log(\omega - 1) - \omega \log(\omega). \tag{22}$$

  This is called the contrast-entropy generator.

Extension: Unknown Propensities

- Regarding the choice of $\omega_i^{(0)}$, we can use

$$\omega_i^{(0)} = \frac{\pi_i^{-1}}{\sum_{i \in S} \pi_i^{-1}}$$

  where π_i is the first-order inclusion probability of unit $i \in S$.
- In practice, π_i is often unknown (non-probability sampling, observational data).
- We assume MAR (i.e. $\delta \perp Y \mid \boldsymbol{x}$) and estimate $\pi(\boldsymbol{x})$ via a flexible learner (logistic regression, GAM, random forest, ...) and set $\hat{\omega}_i^{(0)} \propto \hat{\pi}_i^{-1}$ up to normalization.
- **Problem:** plugging $\hat{\pi}_i$ directly can introduce overfitting bias.
- **Solution:** Cross-fitting (sample splitting).

Cross-Fitting Procedure

Fix $K \geq 2$ folds. For each fold k:

1. Estimate $\hat{\pi}^{(-k)}(\cdot)$ using only data *outside* fold k.
2. Set $\hat{\pi}_i^{(-)} := \hat{\pi}^{(-k)}(\boldsymbol{x}_i)$ for i in fold k.

Cross-fitted baseline weights:

$$\hat{\omega}_i^{(0)} \propto \frac{1}{\hat{\pi}_i^{(-)}}, \quad i \in S.$$

Key benefit: $\hat{\pi}^{(-k)}(\cdot)$ is independent of (y_i, δ_i) in fold k, so the cross-term involving $\varepsilon_i = y_i - m(\boldsymbol{x}_i)$ has conditional mean zero $\Rightarrow$ eliminates overfitting bias, where $m(\boldsymbol{x}) = E(Y \mid X = \boldsymbol{x})$.

Robustness of the Calibration Estimator

- Define the linear approximation error

$$r(\boldsymbol{x}) := m(\boldsymbol{x}) - \boldsymbol{x}^\top \tilde{\boldsymbol{\beta}}_g^{(0)}.$$

  where $\tilde{\boldsymbol{\beta}}_g^{(0)}$ is the probability limit of $\hat{\boldsymbol{\beta}}_g$ in (21).
- **Product-rate condition**:

$$\frac{n}{N} \cdot \sqrt{n}\, \|\hat{d}^{(-)}(\boldsymbol{x}) - d(\boldsymbol{x})\|_{L_2} \cdot \|r(\boldsymbol{x})\|_{L_2} = o_p(1),$$

  where $d(\boldsymbol{x}) = \pi(\boldsymbol{x})^{-1}$.
- This condition is satisfied in two important cases:
  (i) **Outcome-model robustness**: $m(\boldsymbol{x}) = \boldsymbol{x}^\top \tilde{\boldsymbol{\beta}}_g^{(0)}$, so $r(\boldsymbol{x}) \equiv 0 \rightarrow$ holds trivially regardless of how well $\pi(\boldsymbol{x})$ is estimated.
  (ii) **Semiparametric double robustness** at rates n^{-a} and n^{-b}: condition holds whenever $a + b > 1/2$.
- **Key message:** the calibration estimator is $\sqrt{n}$-consistent as long as the product of the two estimation errors vanishes fast enough.

Simulation Study 1: Setup

- $N = 10{,}000$; $x_i = (1, x_{i1}, \ldots, x_{i4})^\top$, $x_{ij} \sim \mathrm{TN}(2, 1, 0, 4)$ iid.
- Two **outcome regression (OR)** models:
  - OR0: $Y_i = 1 + x_{i1} - x_{i2} + e_i$ (linear)
  - OR1: $Y_i = 1 + x_{i1} - x_{i2} + x_{i1}x_{i2} + (x_{i2}^2 - 1) + e_i$ (nonlinear)
- Two **propensity score (PS)** models:
  - PS0: $c_i = -1 - 0.25x_{i2} + 0.5x_{i3}$ (linear logistic)
  - PS1: $c_i = -1 - 0.25(x_{i2} - 3)(x_{i3} - 4) + 0.5(x_{i2} - 2.5)^4$ (nonlinear)
- Four scenarios: PS0/OR0, PS0/OR1, PS1/OR0, PS1/OR1.
- Cross-fitted baseline weights using `glm` or `gam`; $B = 500$ replications.

Simulation Study 1: Estimators

- **IPW**: inverse-probability weighted estimator using $\hat{\pi}_i$.
- **DS**: Deville–Särndal calibration.
- **BC**: Bregman-divergence calibration (proposed).
- Generator choices: ET ($\omega \log \omega$), EL ($-\log \omega$), HD ($(\sqrt{\omega} - 1)^2$).
- Note: DS-ET and BC-ET are *numerically identical* (well-known equivalence).

Doubly robust property: calibration estimators should be consistent when *at least one* of PS or OR is correctly specified.

Simulation Study 1: Results (Bias $\times 10^2$)

Table: MC Bias ($\times 10^2$) and RMSE ($\times 10^2$). ET row = DS-ET = BC-ET.

		PS0/OR0		PS0/OR1		PS1/OR0		PS1/OR1	
$\hat{\pi}$	Method	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
glm	IPW	0.0	1.5	0.0	3.6	−24.0	24.1	180.4	180.6
	ET	0.0	1.4	−0.1	2.2	0.0	1.4	56.4	56.4
	DS-EL	0.0	1.4	−0.1	2.2	0.0	1.4	58.4	58.4
	BC-EL	0.0	1.4	−0.1	2.2	0.0	1.4	49.8	49.8
	BC-HD	0.0	1.4	−0.1	2.2	0.4	1.4	52.9	52.9
gam	IPW	−0.1	1.5	0.1	3.3	−0.8	1.9	4.2	5.0
	ET	0.0	1.4	−0.1	2.0	0.0	1.7	0.2	2.0
	DS-EL	0.0	1.4	−0.1	2.0	0.0	1.7	0.3	2.0
	BC-EL	0.0	1.4	−0.1	2.0	0.0	1.7	0.2	2.0
	BC-HD	0.0	1.4	−0.1	2.0	0.0	1.7	0.2	2.0

Simulation Study 1: Key Findings

- All calibration estimators eliminate IPW bias when *at least one* model is correct (**double robustness confirmed**).
- Under doubly misspecified PS1/OR1 with `glm`:
  - All calibration estimators show substantial bias (linear logistic cannot capture PS1).
  - BC-EL and BC-HD have notably smaller bias than DS counterparts.
- Under doubly misspecified PS1/OR1 with `gam`: all calibration estimators show negligible bias – penalized splines capture the nonlinear PS.
- DS-ET $\equiv$ BC-ET across all configurations (expected from theory).

Simulation Study 2: High-Dimensional Calibration

- $N = 10{,}000$; $p = 500$ covariates, $X_i \sim N(2, \Sigma)$ with $\Sigma_{jk} = 0.5^{|j-k|}$.
- True outcome model: OR0 ($Y_i = 1 + x_{i1} - x_{i2} + e_i$); only 2 relevant covariates.
- Estimators compared:
  - **Full**: exact calibration on all $p = 500$ covariates.
  - **Oracle**: exact calibration on true covariates $(1, X_1, X_2)$ only.
  - **SBC**: soft Bregman calibration with $q = 1, 2, \infty$ and outcome-guided $\tau_k = \tau / |\hat{\beta}_k|$.
- Two pilot estimators for $\hat{\beta}_k$: OLS and LASSO-refit.
- $B = 500$ replications; τ selected by cross-validation.

Simulation Study 2: Results

Table: RMSE ($\times 10^2$) at fixed $\tau = 5 \times 10^{-4}$ for the ET generator.

Pilot	Estimator (q)	Bias ($\times 10^2$)	RMSE ($\times 10^2$)
	IPW	0.528	2.043
	Full	0.004	1.497
	Oracle	-0.021	1.414
OLS	SBC $q = 1$	0.005	1.489
	SBC $q = 2$	0.005	1.474
	SBC $q = \infty$	-0.010	1.417
LASSO-refit	SBC $q = 1$	-0.013	1.416
	SBC $q = 2$	-0.013	1.416
	SBC $q = \infty$	-0.011	1.418

Simulation Study 2: Key Findings

- With OLS pilot, $q = \infty$ (Lasso-type) recovers near-Oracle RMSE; $q = 1, 2$ remain closer to Full.
- With LASSO-refit pilot, *all three q-norms* achieve near-Oracle performance – sparse pilot effectively removes irrelevant covariates.
- Increasing τ beyond threshold reverts estimator toward IPW.
- Choice of Bregman divergence function (ET, EL, HD) has negligible effect on performance in this setting.

Conclusion

- We formulate calibration as a **Bregman projection in weight space** (not on weight ratios), revealing an elegant primal–dual symmetry.
- Duality yields a **low-dimensional convex optimization** in $\boldsymbol{\lambda}$ (p-dim) from an n-dim constrained problem; closed-form recovery of $\hat{\omega}_i$.
- The generator $G(\cdot)$ **tunes the regression coefficient** and hence efficiency; under Poisson sampling, the **contrast-entropy** generator achieves design-optimality.
- With unknown propensities, **cross-fitting** combined with calibration yields estimation requiring only a product-rate condition on model errors.
- For high-dimensional $\boldsymbol{x}$, **soft Bregman calibration** with Hölder-regularized dual achieves near-oracle performance via adaptive variable selection (skipped).

Acknowledgements

- U.S. National Science Foundation (2242820)
- U.S. Department of Agriculture's National Resources Inventory, Cooperative Agreement NR203A750023C006, Great Rivers CESU 68-3A75-18-504.

References I

Anderson, R. L. (1957), 'Maximum likelihood estimates for the multivariate normal distribution when some observations are missing', *Journal of the American Statistical Association* **52**, 200–203.

Cho, Geonhee, Jae Kwang Kim and Yumou Qiu (2026), 'A tutorial on bregman projection in statistics'. Unpublished manuscript.

Deville, J. C. and C. E. Särndal (1992), 'Calibration estimators in survey sampling', *Journal of the American Statistical Association* **87**, 376–382.

Fuller, W. A. (1975), 'Regression analysis for sample survey', *Sankhya C* **17**, 117–132.

Hirshberg, David A, Arian Maleki and Jose R Zubizarreta (2021), 'Minimax linear estimation of the retargeted mean', *arXiv preprint arXiv:1901.10296* .

Kim, Jae Kwang, Yonghyun Kwon and Yumou Qiu (2026), 'Bregman projection for calibration estimation'. Submitted (arXiv:2603.20780)).

References II

Montanari, G. E. (1987), 'Post-sampling efficient qr-prediction in large-scale surveys', *International Statistical Review* **55**, 191–202.