Fairness Constrained Sampling

Ananyapam De

TU Clausthal

November 6, 2025

Example

- Task: predict if income $> \$50\text{k}$ from socio-economic features.
- Binary label $y \in \{0, 1\}$ and features $x \in \mathbb{R}^d$.
- Goal: be *accurate* and *fair* across groups.
- Dataset size: 32,561 rows, 14 features.

Age	Sex	Education	Occupation	Income
39	Male	Bachelors	Adm-clerical	¡=50K
50	Male	Bachelors	Exec-managerial	¡=50K
38	Male	HS-grad	Handlers-cleaners	¡=50K
52	Male	HS-grad	Exec-managerial	¿50K
53	Male	11th	Handlers-cleaners	¡=50K
⋮	⋮	⋮	⋮	⋮

Table: Adult dataset (UCI)

Bayesian Logistic Regression

Model

$$y_n \mid x_n, \theta \sim \text{Bernoulli}(p_n),$$
$$p_n = \sigma(x_n^\top \theta) = \frac{1}{1 + e^{-x_n^\top \theta}}.$$

- **Likelihood:** $p(\mathcal{D} \mid \theta) = \prod_{n=1}^{N} [\sigma(x_n^\top \theta)]^{y_n} [1 - \sigma(x_n^\top \theta)]^{1-y_n}$.

Bayesian Logistic Regression

Model

$$y_n \mid x_n, \theta \sim \text{Bernoulli}(p_n),$$
$$p_n = \sigma(x_n^\top \theta) = \frac{1}{1 + e^{-x_n^\top \theta}}.$$

- **Likelihood:** $p(\mathcal{D} \mid \theta) = \prod_{n=1}^{N} [\sigma(x_n^\top \theta)]^{y_n} [1 - \sigma(x_n^\top \theta)]^{1-y_n}$.
- **Prior:** $p(\theta) = \mathcal{N}(0, \sigma^2 I)$.

Bayesian Logistic Regression

Model

$$y_n \mid x_n, \theta \sim \text{Bernoulli}(p_n),$$
$$p_n = \sigma(x_n^\top \theta) = \frac{1}{1 + e^{-x_n^\top \theta}}.$$

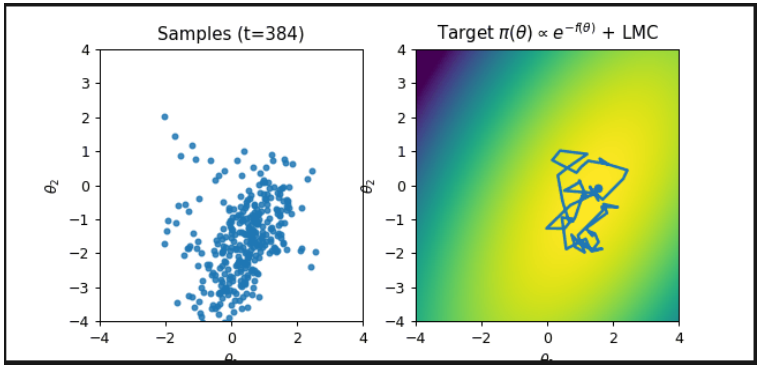
- **Likelihood:** $p(\mathcal{D} \mid \theta) = \prod_{n=1}^{N} [\sigma(x_n^\top \theta)]^{y_n} [1 - \sigma(x_n^\top \theta)]^{1-y_n}$.
- **Prior:** $p(\theta) = \mathcal{N}(0, \sigma^2 I)$.
- **Posterior:** $\pi(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D}|\theta)\, p(\theta)}{p(\mathcal{D})} \propto \exp(\underbrace{\log p(\mathcal{D} \mid \theta) + \log p(\theta)}_{-U(\theta)})$.

Langevin Monte Carlo

Discretization (LMC)

Euler-Maruyama discretization with step size η:

$$\theta_{k+1} = \theta_k - \eta \nabla U(\theta_k) + \sqrt{2\eta}\,\xi_k$$



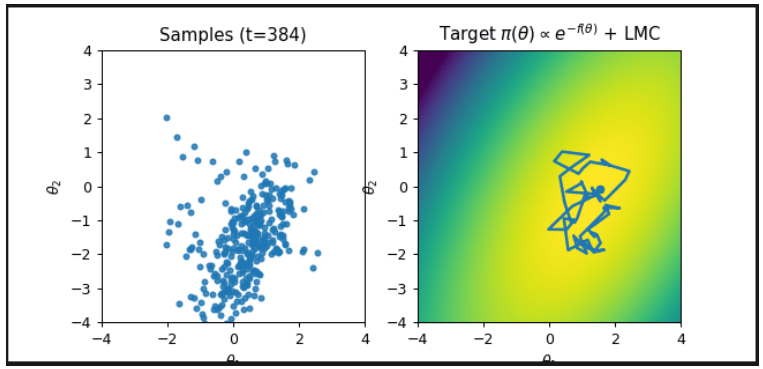
Langevin Monte Carlo

Discretization (LMC)

Euler-Maruyama discretization with step size η:

$$\theta_{k+1} = \theta_k - \eta \nabla U(\theta_k) + \sqrt{2\eta}\,\xi_k$$

- Gradient term: *drift* toward high-probability regions.



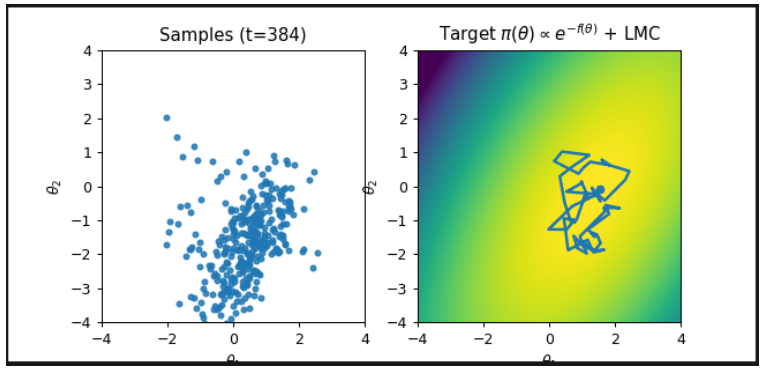
Langevin Monte Carlo

Discretization (LMC)

Euler-Maruyama discretization with step size η:

$$\theta_{k+1} = \theta_k - \eta \nabla U(\theta_k) + \sqrt{2\eta}\, \xi_k$$

- Gradient term: *drift* toward high-probability regions.
- Noise term: *exploration* to avoid getting stuck, where $\xi_k \sim \mathcal{N}(0, I)$.



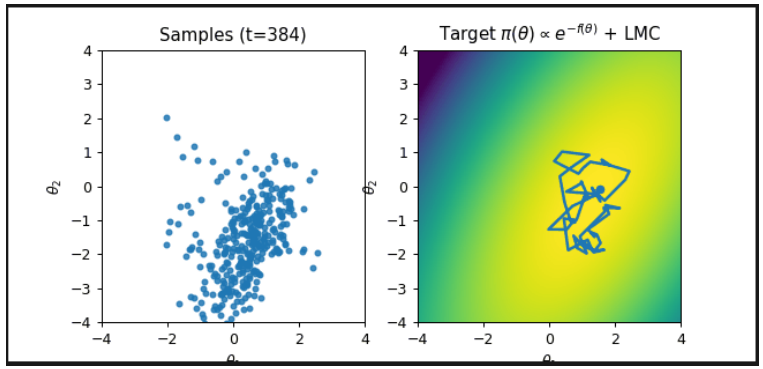
Langevin Monte Carlo

Discretization (LMC)

Euler-Maruyama discretization with step size η:

$$\theta_{k+1} = \theta_k - \eta \nabla U(\theta_k) + \sqrt{2\eta}\, \xi_k$$

- Gradient term: *drift* toward high-probability regions.
- Noise term: *exploration* to avoid getting stuck, where $\xi_k \sim \mathcal{N}(0, I)$.
- As $\eta \to 0$, the chain's stationary distribution approaches $\pi(\theta \mid \mathcal{D})$.



The World is Biased, Sadly

The World is Biased, Sadly

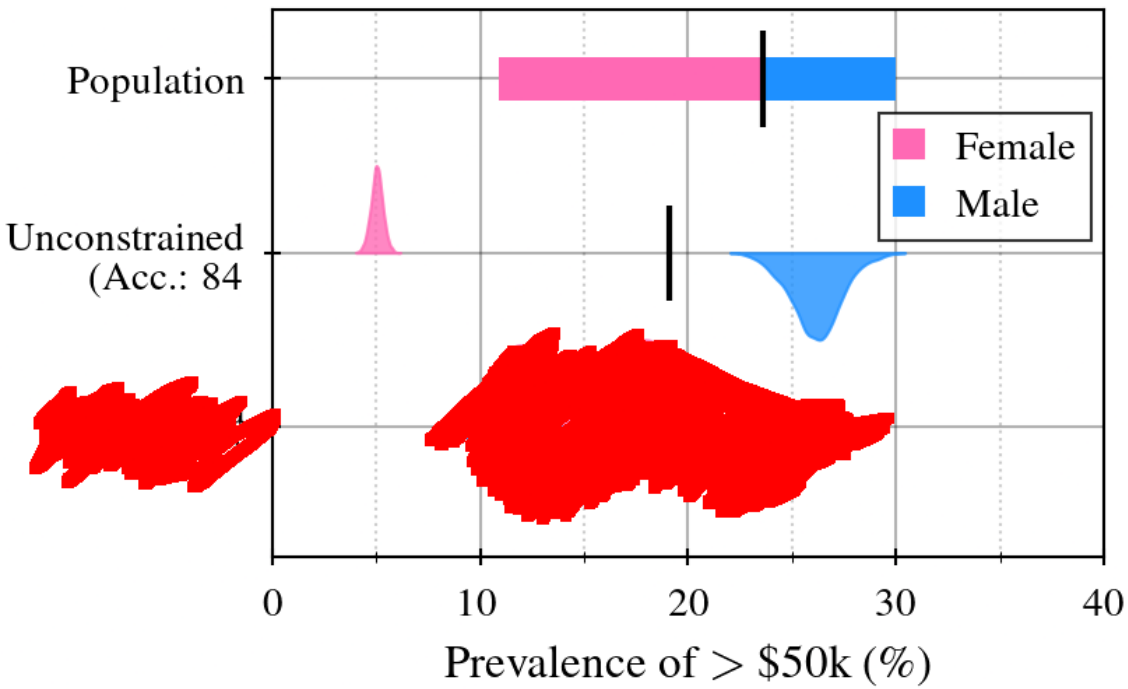
- Under the posterior, average probability of $> 50\text{k}$ is:
  - Overall: **19.1%** Male: **26.2%** Female: **5.0%**.

The World is Biased, Sadly

- Under the posterior, average probability of $> 50\text{k}$ is:
  - Overall: **19.1%** Male: **26.2%** Female: **5.0%**.
- The posterior *amplifies* disparity across genders.

The World is Biased, Sadly

- Under the posterior, average probability of $> 50\text{k}$ is:
  - Overall: **19.1%** Male: **26.2%** Female: **5.0%**.
- The posterior *amplifies* disparity across genders.



Notion of Fairness: Statistical parity

Notion of Fairness: Statistical parity

What it measures

Do different groups receive *positive predictions* at the same rate?

Notion of Fairness: Statistical parity

What it measures

Do different groups receive *positive predictions* at the same rate?

Notation

- Groups: $G \in \{A, B\}$
- For person n with features x_n, model outputs: $\hat{p}_n = \sigma(x_n^\top \theta)$
- Rate in group G: $\text{Rate}_G(\theta) = \frac{1}{|G|} \sum_{n \in G} \hat{p}_n$

Notion of Fairness: Statistical parity

What it measures

Do different groups receive *positive predictions* at the same rate?

Notation

- Groups: $G \in \{A, B\}$
- For person n with features x_n, model outputs: $\hat{p}_n = \sigma(x_n^\top \theta)$
- Rate in group G: $\text{Rate}_G(\theta) = \frac{1}{|G|} \sum_{n \in G} \hat{p}_n$

Statistical parity

$$\text{Rate}_A(\theta) = \text{Rate}_B(\theta)$$

(Often relaxed to: $|\text{Rate}_A(\theta) - \text{Rate}_B(\theta)| \leq \delta$)

Constraining the distribution

Optimization Problem

$$\min_{\mu} \ \mathrm{KL}(\mu \,\|\, \pi) \quad \text{s.t.} \quad \mathbb{E}_{\theta \sim \mu}[g(\theta)] \leq 0.$$

- μ: The constrained distribution over parameters.
- π: The original Bayesian posterior.
- $g(\theta)$: Constraint functions encoding fairness requirements.
- Minimizing KL divergence keeps μ close to π while enforcing fairness.

Tilting the posterior

- Lagrangian over measures:

$$\mathcal{L}(\mu, \lambda) = \mathrm{KL}(\mu \,\|\, \pi) + \lambda^\top \mathbb{E}_{\theta \sim \mu}[g(\theta)], \qquad \lambda \geq 0.$$

$$\boxed{\mu(\theta) \propto \pi(\theta) \exp\big\{ -\lambda^\top g(\theta) \big\}}$$

- **Interpretation:** The exponential term *reweights* the posterior.

Tilting the posterior

- Lagrangian over measures:

$$\mathcal{L}(\mu, \lambda) = \mathrm{KL}(\mu \,\|\, \pi) + \lambda^\top \mathbb{E}_{\theta\sim\mu}[g(\theta)], \qquad \lambda \geq 0.$$

$$\boxed{\mu(\theta) \propto \pi(\theta) \exp\big\{ -\lambda^\top g(\theta)\big\}}$$

- **Interpretation:** The exponential term *reweights* the posterior.
  - If $g(\theta) > 0$ (constraint violated), $\exp\{-\lambda^\top g(\theta)\} < 1 \Rightarrow$ *downweight* θ.

Tilting the posterior

- Lagrangian over measures:

$$\mathcal{L}(\mu, \lambda) = \mathrm{KL}(\mu \,\|\, \pi) + \lambda^\top \mathbb{E}_{\theta\sim\mu}[g(\theta)], \qquad \lambda \geq 0.$$

$$\boxed{\mu(\theta) \propto \pi(\theta) \exp\big\{ -\lambda^\top g(\theta)\big\}}$$

- **Interpretation:** The exponential term *reweights* the posterior.
  - If $g(\theta) > 0$ (constraint violated), $\exp\{-\lambda^\top g(\theta)\} < 1 \Rightarrow$ *downweight* θ.
  - If $g(\theta) \leq 0$ (constraint satisfied), $\exp\{-\lambda^\top g(\theta)\} \geq 1 \Rightarrow$ *upweight* θ.

Tilting the posterior

- Lagrangian over measures:

$$\mathcal{L}(\mu, \lambda) = \mathrm{KL}(\mu \,\|\, \pi) + \lambda^\top \mathbb{E}_{\theta\sim\mu}[g(\theta)], \qquad \lambda \geq 0.$$

$$\boxed{\mu(\theta) \propto \pi(\theta) \exp\big\{ -\lambda^\top g(\theta)\big\}}$$

- **Interpretation:** The exponential term *reweights* the posterior.
  - If $g(\theta) > 0$ (constraint violated), $\exp\{-\lambda^\top g(\theta)\} < 1 \Rightarrow$ *downweight* θ.
  - If $g(\theta) \leq 0$ (constraint satisfied), $\exp\{-\lambda^\top g(\theta)\} \geq 1 \Rightarrow$ *upweight* θ.
  - Larger $\lambda \Rightarrow$ stronger penalty for violations.

Tilting the posterior

- Lagrangian over measures:

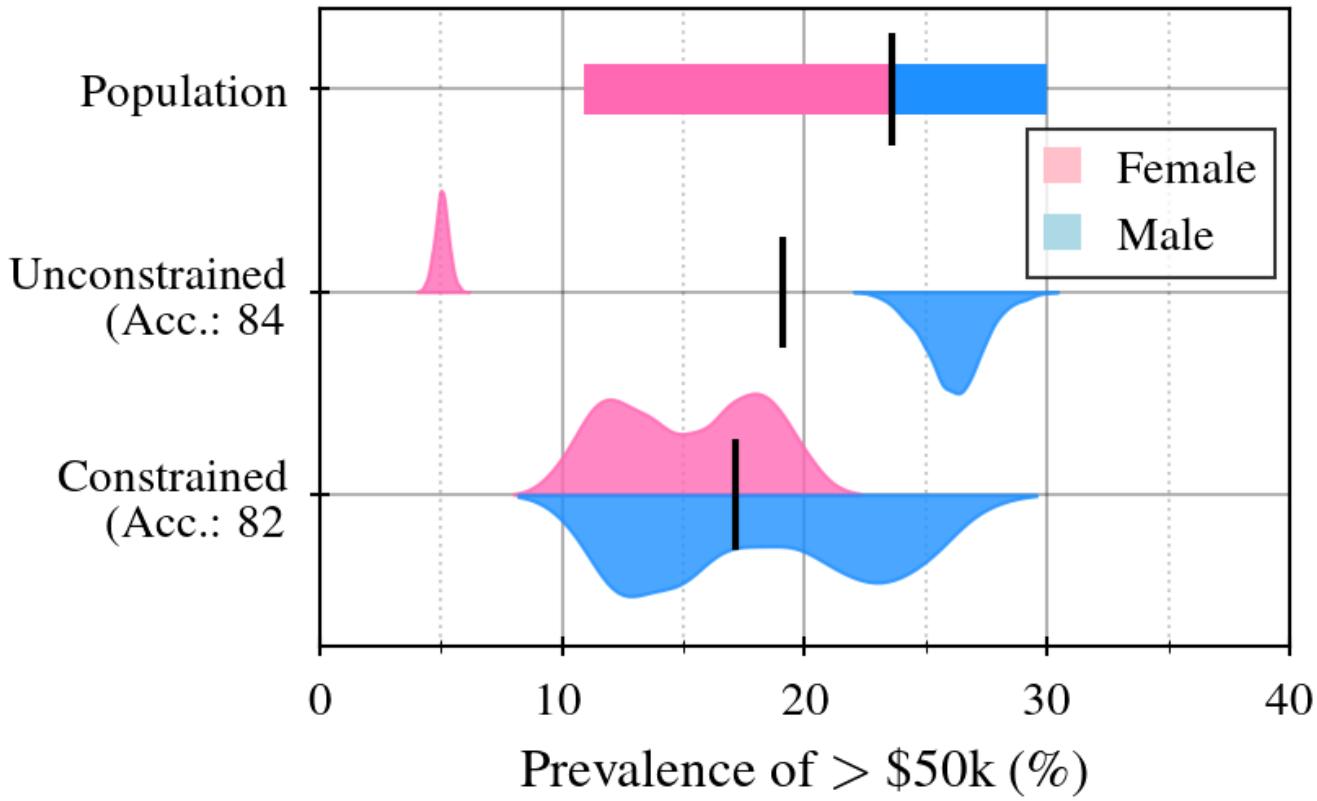
$$\mathcal{L}(\mu, \lambda) = \mathrm{KL}(\mu \,\|\, \pi) + \lambda^\top \mathbb{E}_{\theta\sim\mu}[g(\theta)], \qquad \lambda \geq 0.$$

$$\boxed{\mu(\theta) \propto \pi(\theta) \exp\big\{ -\lambda^\top g(\theta)\big\}}$$

- **Interpretation:** The exponential term *reweights* the posterior.
  - If $g(\theta) > 0$ (constraint violated), $\exp\{-\lambda^\top g(\theta)\} < 1 \Rightarrow$ *downweight* θ.
  - If $g(\theta) \leq 0$ (constraint satisfied), $\exp\{-\lambda^\top g(\theta)\} \geq 1 \Rightarrow$ *upweight* θ.
  - Larger $\lambda \Rightarrow$ stronger penalty for violations.
  - Tilting *shifts probability mass* toward parameters that satisfy statistical parity.

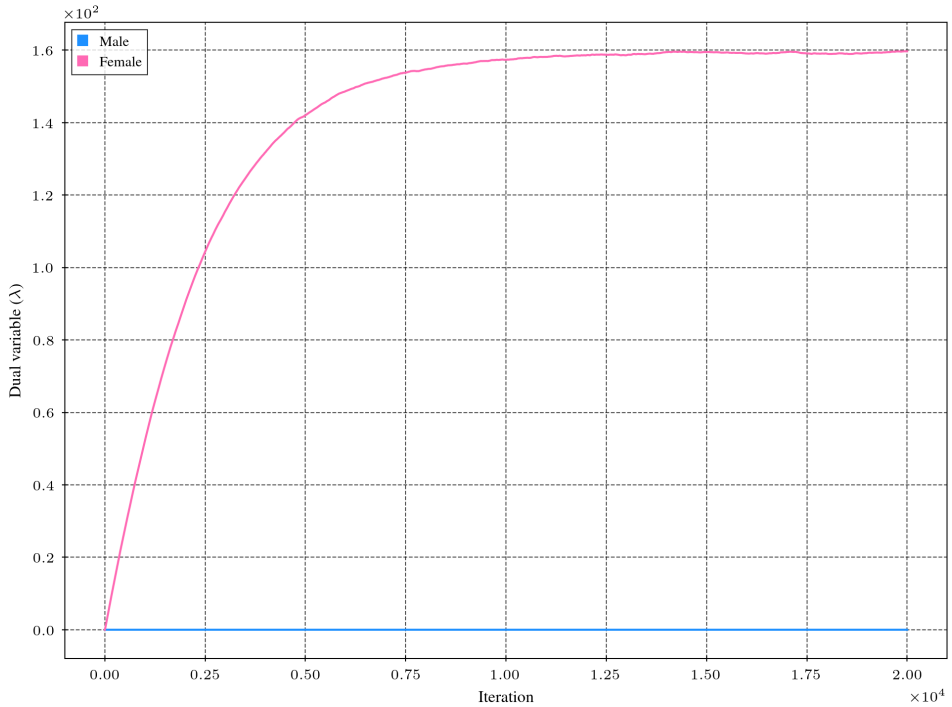
Adult dataset: what changes under the constraint?

- Unconstrained posterior: overall **19.1%**, male **26.2%**, female **5.0%**.
- Constrained (e.g., $\delta = 0.01$): overall **17.1%**, male **18.1%**, female **15.1%**.
- Accuracy: slight drop (84% $\rightarrow$ 82%), but disparity is greatly reduced.

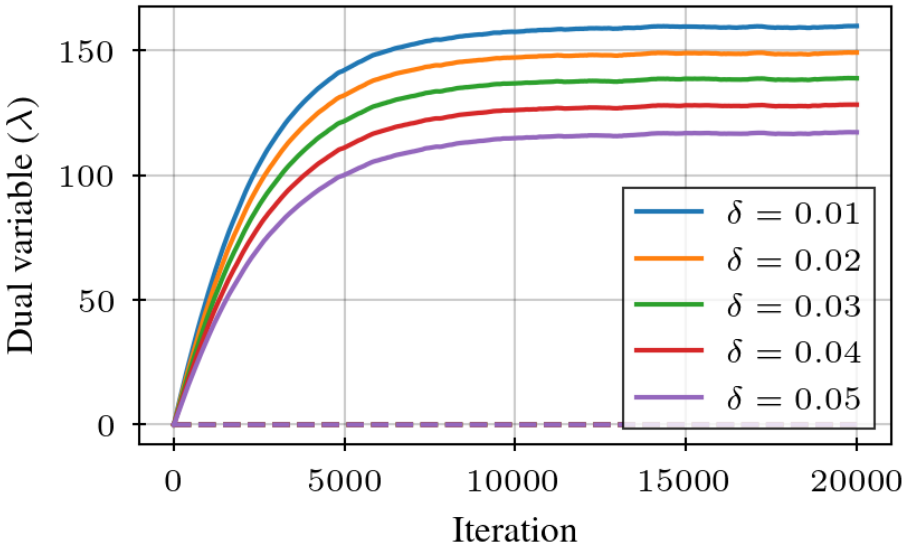


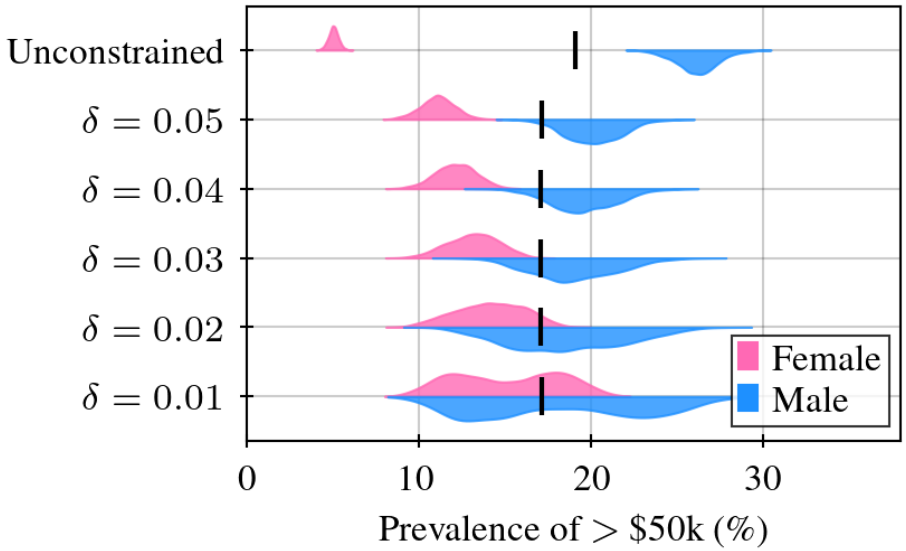
Interpreting the dual: which constraint binds?

- In the experiment, the *female* constraint is binding (large λ); the male constraint stays inactive ($\lambda \approx 0$).
- Duals act like "*explainers*": they show *which* fairness requirement shapes the final samples.



Relaxing Constraints





- As the constraint tolerance δ increases, the distributions converge toward the unconstrained posterior.
- Tighter constraints ($\delta \to 0$) enforce stronger fairness, while relaxed constraints allow more disparity.

Takeaways

Takeaways

- Bayesian Inference deals with distributions of parameters, NOT just parameters.

Takeaways

- Bayesian Inference deals with distributions of parameters, NOT just parameters.
- If we learn a model from biased data, we get a biased model.

Takeaways

- Bayesian Inference deals with distributions of parameters, NOT just parameters.
- If we learn a model from biased data, we get a biased model.
- Instead of trying to unbias the data, we tilt the posterior distribution by constraining.

Takeaways

- Bayesian Inference deals with distributions of parameters, NOT just parameters.
- If we learn a model from biased data, we get a biased model.
- Instead of trying to unbias the data, we tilt the posterior distribution by constraining.
- Small accuracy tradeoff for large disparity reduction.

Ideas

Currently figuring out:

Ideas

Currently figuring out:

- Distributional fairness

Ideas

Currently figuring out:

- Distributional fairness
- Augmentation to make the optimization more smoother

Ideas

Currently figuring out:

- Distributional fairness
- Augmentation to make the optimization more smoother

Thank you!