Augmented Lagrangian Langevin Monte Carlo

Ananyapam De Benjamin Säfken

Institut für Mathematik, TU Clausthal

TU Clausthal
Clausthal University of Technology

github.com/Ananya-pam7/allmc

Problem of Fair Sampling

Consider data pairs (x, y) with $x \in \mathcal{X}$, $y \in \{0, 1\}$, and protected subgroups $\mathcal{G}_i \subset \mathcal{X}$, $i \in I$. Let π be a Bayesian posterior over parameters $\theta \in \mathbb{R}^d$ and $q(x; \theta) \in [0, 1]$ the model's positive-class probability. We wish to enforce statistical parity such that

$$\mathbb{E}_{x,\,\theta\sim\pi}[q(x;\theta) \mid \mathcal{G}_i] \geq \mathbb{E}_{x,\,\theta\sim\pi}[q(x;\theta)] - \delta_i.$$

Compactly, the constraint can be written as

$$\mathbb{E}_{\theta\sim\mu}\big[g_i(\theta)\big] \leq 0,$$

$$\text{where } p_i = \rho_{\mathcal{X}}(\mathcal{G}_i), \quad g_i(\theta) = \mathbb{E}_{x\sim\rho_{\mathcal{X}}}\big[p_i q(x;\theta) - q(x;\theta)\mathbf{1}_{x\in\mathcal{G}_i} - p_i\delta_i\big].$$

But π itself may not satisfy the constraint!

Primal- Dual Problem Setup

Constrained sampling (primal):

$$P^\star = \min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathrm{KL}(\mu\|\pi) \quad \text{s.t.} \quad \mathbb{E}_{x\sim\mu}[g_i(x)] \leq 0,\ i = 1, \ldots, I.$$

Here $g(\cdot)$ encodes statistical constraints. We solve this using a Lagrangian approach with augmentation.

$$\mathcal{L}_\rho(\mu, \lambda) = \mathrm{KL}(\mu\|\pi) + \lambda^\top \mathbb{E}_\mu[g] + \frac{\rho}{2}\left\|\mathbb{E}_\mu[g]\right\|^2, \qquad \lambda \geq 0,\ \rho \geq 0.$$

Introducing a **tracker** $c \approx \mathbb{E}_\mu[g]$ we form the surrogate pointwise potential

$$U_\rho(x, \lambda, c) = f(x) + \lambda^\top g(x) + \frac{\rho}{2}\|g(x) - c\|^2,$$

so that $\mu_{\lambda,c} \propto e^{-U_\rho(\cdot,\lambda,c)}$ is tractable for LMC. We define the dual function

$$d_\rho(\lambda) \triangleq \inf_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{L}_\rho(\mu, \lambda).$$

$d_\rho(\lambda)$ is a concave function of λ and yields the lower bound $d_\rho(\lambda) \leq P^\star$ for all $\lambda \geq 0$. The dual problem is therefore

$$D^\star_\rho := \max_{\lambda\in\mathbb{R}^I_+} d_\rho(\lambda),$$

and the maximizers $\Phi^\star = \mathrm{argmax}_{\lambda\geq 0}\, d_\rho(\lambda)$ play the role of (regularized) Lagrange multipliers.

Gradient Flow Structure

This structure suggests a primal dual tracker gradient flow in the Wasserstein space: we simultaneously update the distribution μ, the multipliers λ, and the tracker c according to

$$\frac{\partial\mu(t)}{\partial t} = \nabla\cdot\big(\mu(t)\,\nabla_{\mathcal{W}_2}\mathcal{L}_\rho(\mu(t), \lambda(t))\big), \tag{1}$$

$$\frac{\partial\lambda_i(t)}{\partial t} = \big[\nabla_{\lambda_i}\mathcal{L}_\rho(\mu(t), \lambda(t))\big]_{\lambda_i(t),+} = \big[\mathbb{E}_{\mu(t)}[g_i]\big]_{\lambda_i(t),+}, \tag{2}$$

$$\dot c(t) = \kappa\big(\mathbb{E}_{\mu(t)}[g] - c(t)\big), \qquad \kappa > 0, \tag{3}$$

where $[z]_{\lambda,+} = \max(z, 0)$. The c-dynamics forces the surrogate center to follow the current constraint expectation so that the pointwise potential U_ρ faithfully approximates the quadratic penalty. We discretize the primal flow (1) via Euler–Maruyama on the surrogate potential U_ρ and implement the dual and tracker ODEs (2)–(3) with stochastic, single-sample updates.

Algorithm

- **Primal step (Langevin):** Euler–Maruyama on U_ρ:

$$x_{k+1} = x_k - \eta_k\nabla_x U_\rho(x_k, \lambda_k, c_k) + \sqrt{2\eta_k}\,\xi_k, \quad \xi_k \sim \mathcal{N}(0, I_d)$$

- **Tracker (EMA):**

$$c_{k+1} = (1 - \beta_k)c_k + \beta_k\, g(x_{k+1}).$$

- **Dual (ascent):**

$$\lambda_{k+1} = \max\{0,\ \lambda_k + \alpha_k\, g(x_{k+1})\}.$$

Practical note: Compute $g(x)$ as the data-averaged at parameter x, i.e. $\bar g(x) = \mathbb{E}_{z\sim\rho_X}[\bar g(z, x)]$.

Key theoretical properties (informal)

- **Convergence.** Under the same smoothness / ergodicity assumptions used for PD-LMC, AL-LMC attains the same averaged KL / Wasserstein rates up to a controllable tracking bias; with two-timescale choices the bias vanishes asymptotically.
- **Tracking error (expectation).** Let $m_k = \mathbb{E}_{\mu_{\lambda_k,c_k}}[g]$ and $\Delta_k = c_k - m_k$. For constant steps (η, α, β) one has the expectation bound

$$\limsup_{k\to\infty}\mathbb{E}\|\Delta_k\| \;=\; O\Big(\frac{\rho_{\mathrm{mix}} + \eta}{\beta} + \frac{\alpha}{\beta} + \beta\Big),$$

where ρ_{mix} is the one-step mixing bias of the inner LMC kernel.

Interpretation. the steady-state error has three sources: (i) sampling/mixing error $\varepsilon_{\mathrm{mix}}(m)$ divided by β (so run more inner steps m or increase β); (ii) movement of the dual α relative to β (so choose $\alpha \ll \beta$); (iii) intrinsic tracker noise $\sim \beta$. Typical practical rule: choose $\alpha \ll \beta$ and pick m (or β) so that $\varepsilon_{\mathrm{mix}}(m) \lesssim \beta$.

- **Tracking error (concentration).** for any fixed horizon and confidence δ there is a matching high-probability bound of the form

$$\|\Delta_k\| = O\Big(\frac{\rho_{\mathrm{mix}} + \eta}{\beta} + \frac{\alpha}{\beta} + \beta + \sqrt{\beta\log(1/\delta)}\Big)$$

with probability at least $1 - \delta/2$.

Practical consequence. Choose a time-scale separation $\alpha \ll \beta$ and set β so that $\varepsilon_{\mathrm{mix}}(m) \lesssim \beta$ (or run m inner LMC steps) to make the tracking bias negligible.

Lagrangian vs Augmented Lagrangian

Metric	PD-LMC (Lagrangian)	AL-LMC (Augmented)
Max stable dual step $\alpha_{\max}$	0.20	**0.45**
Steady-state dual variance (norm)	1.00	**0.85**
Primal mixing time (effective)	1.0 (normalized)	1.3 (increases with ρ)
Constraint satisfaction (bias)	0.02	**0.005**

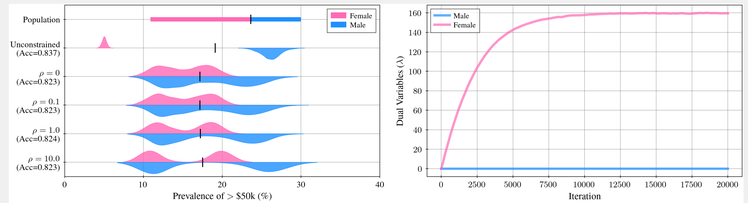
Experiments (Adult dataset)

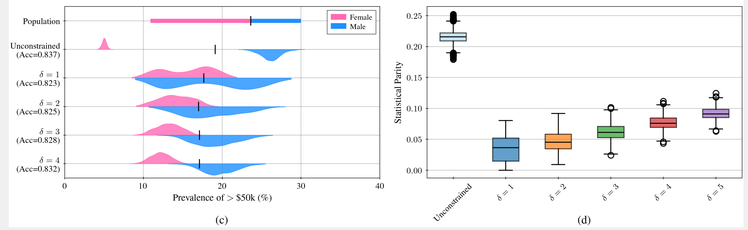
Setup. We train a Bayesian logistic regression on the Adult dataset ($N = 32,561$, $d = 62$) to predict income $>$ \$50k. The posterior uses a binomial log-likelihood with independent zero-mean Gaussian priors ($\sigma^2 = 3$). Gender is treated as the protected attribute with two groups $\mathcal{G}_{\text{male}}, \mathcal{G}_{\text{female}}$ ($I = 2$).

Fairness constraint. We enforce statistical parity via the empirical constraint

$$\bar g_i(\theta) = \frac{1}{|\mathcal{G}_i|}\sum_{n\in\mathcal{G}_i} q(x_n;\theta) - \frac{1}{N}\sum_{n=1}^{N} q(x_n;\theta) - \delta_i,$$

with tolerance $\delta_i = 0.01$.





Results (summary). AL-LMC substantially reduces gender disparity while preserving accuracy. Unconstrained LMC (baseline) — the unconstrained posterior exhibits strong gender disparity: prevalences of male 26.2% vs female 0.05%, amplifying test-set prevalences: male 30.6% vs female 10.9%. With parity constraints the constrained posterior yields prevalences: overall 17.1% male 18.1%,female 15.1%. This reduces disparity with only a small drop in accuracy (from 84% to 82%).

Practical tuning recommendations

- Choose ρ moderately: too small $\Rightarrow$ weak dual regularization; too large $\Rightarrow$ primal stiffness. Start with $\rho \in [0.1, 1]$.
- Two-timescale: $\alpha \ll \beta \ll 1$. E.g. $\beta \asymp (\varepsilon_{\mathrm{mix}} + \eta)^{1/2}$, $\alpha \asymp \beta/10$.
- If primal mixing is slow, run several LMC steps per dual/tracker update (reduces $\varepsilon_{\mathrm{mix}}$).

References and further reading

- L. F. O. Chamon, M. R. Karimi, A. Korba, "Constrained Sampling with Primal–Dual Langevin Monte Carlo", NeurIPS 2024; arXiv:2411.00568.
- M. Welling and Y. W. Teh, "Bayesian Learning via Stochastic Gradient Langevin Dynamics", ICML 2011.
- A. S. Dalalyan, "Theoretical guarantees for approximate sampling from smooth and log-concave densities", J. R. Stat. Soc. Ser. B (2017); (arXiv:1412.7392).
- A. Durmus and É. Moulines, "Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm", Ann. Appl. Probab. (2017); (arXiv:1507.05021).
- Augmented Lagrangian / method of multipliers — classic references: M. R. Hestenes, "Multiplier and Gradient Methods" (1969); M. J. D. Powell, "A Method for Nonlinear Constraints" (1969); D. P. Bertsekas, "Constrained Optimization and Lagrange Multiplier Methods" (Academic Press, 1982; later editions).