*J. Math. Math. Sci. (2025) nn–nnn*

# Adaptive projection-free methods for constrained variational inequalities in machine learning

**Pham Thanh Hieu** (Hanoi, Vietnam)

(Received Sept. 17, 2025; accepted Oct. 8, 2025)

**Abstract.** We propose the *Penalty-Regularized Adaptive Constrained Gradient Method* (PR-A-CGM), a projection-free algorithm for solving variational inequality problems with pseudo-monotone operators and convex functional constraints. Unlike projection-based methods, PR-A-CGM enforces feasibility by introducing a smooth penalty term into the update direction, avoiding costly or intractable projections while retaining convergence guarantees under pseudo-monotonicity and noisy oracles. We prove weak convergence under standard assumptions, establish strong convergence and rates under stronger conditions, and validate our method on machine learning applications such as fairness-constrained classification. Experiments show that PR-A-CGM improves feasibility and robustness over projection-free baselines while narrowing the gap to projection-based methods. These results highlight penalty-regularized primal methods as practical tools for constrained optimization in modern large-scale learning.

## 1. Introduction

Variational inequality problems (VIPs) provide a unifying framework for constrained optimization, game theory, and equilibrium analysis [8]. Given a closed convex set $C \subseteq \mathcal{H}$ and an operator $F : \mathcal{H} \to \mathcal{H}$, a solution $x^* \in C$ satisfies

$$\langle F(x^*), x - x^* \rangle \geq 0 \quad \forall x \in C.$$

*Key words and phrases*: Adaptive Constrained Gradient Method, Pseudo-Monotone, Variational Inequalities
*2020 Mathematics Subject Classification*: 49J40, 90C33, 47H17.

Classical methods such as projected gradient descent (PGD) [18], the extragradient method [12], and the proximal point algorithm [21] guarantee convergence when $F$ is monotone and projections onto $C$ are efficient.

Modern machine learning settings often violate these assumptions. First, operators are frequently only *pseudo-monotone*, arising in nonconvex-concave min-max optimization and adversarial training [5, 14, 9]. Second, feasible sets are defined by functional or non-separable constraints—such as fairness metrics or statistical divergences—where projections are computationally prohibitive or unavailable [17, 4, 6]. In such cases, projection-based schemes (PGD, SEG, PDHG) [3, 10] either become intractable or unstable.

To mitigate this, Zhang et al. [4] proposed the Constrained Gradient Method (CGM), a projection-free, primal method. While effective under monotonicity and exact oracle access, CGM fails in pseudo-monotone and noisy regimes, leading to instability and persistent constraint violations.

Our contribution. We introduce the **Penalty-Regularized Adaptive Constrained Gradient Method (PR-A-CGM)**, a projection-free algorithm that augments CGM with a smooth penalty mechanism and adaptive step sizes. Our main contributions are:

- A novel penalized update rule that enforces feasibility without projections or dual variables,
- Convergence guarantees: weak convergence under pseudo-monotonicity, strong convergence and rates under stronger assumptions,
- Empirical validation on fairness-aware classification, showing improved feasibility and robustness over projection-free baselines and competitive performance relative to projection-based methods.

Together, these results position PR-A-CGM as a scalable alternative for constrained pseudo-monotone optimization in noisy, high-dimensional machine learning environments.

## 2. Preliminaries

We begin by formalizing the setting of constrained variational inequalities in Hilbert spaces and introducing the standing assumptions and technical definitions used throughout the paper.

Let $\mathcal{H}$ be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|x\| := \sqrt{\langle x, x \rangle}$. We consider optimization and equilibrium problems constrained to a feasible set $C \subset \mathcal{H}$ defined by functional inequalities:

$$C := \{x \in \mathcal{H} \mid g_i(x) \leq 0,\ i = 1, \ldots, m\},$$

where each constraint function $g_i : \mathcal{H} \to \mathbb{R}$ is convex, Fréchet differentiable, and has Lipschitz continuous gradients. This general formulation encompasses structural constraints common in applications, such as norm bounds, fairness metrics, and divergence constraints.

Given a continuous operator $F : \mathcal{H} \to \mathcal{H}$, the (Stampacchia-type) variational inequality problem (VI) is to find $x^* \in C$ such that

$$\langle F(x^*), x - x^* \rangle \geq 0 \quad \forall x \in C.$$

This framework generalizes convex optimization, saddle-point formulations, and equilibrium problems. Unlike much of the classical literature, we do not assume monotonicity of $F$, but instead work with the weaker notion of pseudo-monotonicity.

**Definition 2.1** (Pseudo-Monotonicity)**.** *An operator $F : \mathcal{H} \to \mathcal{H}$ is* pseudo-monotone *if, for all $x, y \in \mathcal{H}$,*

$$\langle F(x), y - x \rangle \geq 0 \ \Rightarrow \ \langle F(y), y - x \rangle \geq 0.$$

Pseudo-monotonicity includes monotone operators as a special case and arises naturally in min-max learning problems, adversarial training (e.g., GANs), and nonlinear Nash equilibrium models. However, it complicates algorithmic analysis since uniqueness of solutions and standard monotone convergence guarantees no longer hold.

In practice, $F$ is often available only through a stochastic or noisy oracle. At each iteration $t$, we observe

$$\tilde{F}(x_t) = F(x_t) + \delta_t,$$

where the noise satisfies $\|\delta_t\| \leq \epsilon_t$ and $\sum_{t=0}^{\infty} \epsilon_t < \infty$. This bounded-error model reflects stochastic gradients from mini-batch sampling, simulation errors, or imperfect feedback, and is standard in large-scale learning and online optimization.

For the convergence analysis, we adopt the following standing assumptions:

**(A1) Feasibility:** The constraint set $C$ is nonempty, closed, and convex. Either $C$ is bounded or $F$ is coercive on $C$.

**(A2) Operator:** $F$ is continuous and pseudo-monotone.

**(A3) Constraint Regularity:** Each $g_i$ is convex, Fréchet differentiable, and has $L_i$-Lipschitz continuous gradient:

$$\|\nabla g_i(x) - \nabla g_i(y)\| \leq L_i\|x - y\|, \quad \forall x, y \in \mathcal{H}.$$

**(A4) Oracle Inexactness:** The noisy oracle satisfies $\tilde{F}(x_t) = F(x_t) + \delta_t$, with $\|\delta_t\| \leq \epsilon_t$ and $\sum_{t=0}^{\infty} \epsilon_t < \infty$.

**(A5) Penalty Schedule:** The penalty sequence $\lambda_t$ satisfies $\lambda_t \nearrow \infty$ and $\lambda_t = o(1/\eta_t)$, ensuring feasibility is enforced asymptotically without dominating descent.

For clarity, we recall several technical notions used in later analysis.

**Definition 2.2** (Weak Accumulation Point)**.** *A point $x^\star \in \mathcal{H}$ is a* weak sequential accumulation point *of a sequence $(x_t)$ if there exists a subsequence $(x_{t_k})$ such that $x_{t_k} \rightharpoonup x^\star$ in the weak topology of $\mathcal{H}$.*

**Definition 2.3** (Gap Function)**.** *For $x \in C$, the variational inequality gap is defined as*

$$Gap(x) := \sup_{y \in C} \langle F(x), x - y \rangle.$$

*This function is always nonnegative and vanishes if and only if $x$ solves the VI.*

**Definition 2.4** (Strong Pseudo-Monotonicity)**.** *An operator $F : \mathcal{H} \to \mathcal{H}$ is $\mu$-strongly pseudo-monotone ($\mu > 0$) if, for every $x \in C$ and the solution $x^* \in C$,*

$$\langle F(x), x - x^* \rangle \geq \mu\|x - x^*\|^2.$$

**Definition 2.5** (Lyapunov Function)**.** *Given a solution $x^* \in C$, we define the Lyapunov function*

$$V_t := \tfrac{1}{2}\|x_t - x^*\|^2,$$

*which serves as a merit function to track convergence of the iterates.*

Our objective is to design a fully primal, projection-free algorithm that computes a weak solution of the VI under pseudo-monotonicity and oracle inexactness. The proposed PR-A-CGM algorithm achieves this by incorporating a smooth penalty mechanism, which enforces feasibility asymptotically while ensuring robust convergence guarantees in Hilbert space.

## 3. The PR-A-CGM Algorithm

We introduce the *Penalty-Regularized Adaptive Constrained Gradient Method* (PR-A-CGM), a fully primal, projection-free algorithm designed to solve variational inequality problems with pseudo-monotone operators and functional

inequality constraints. PR-A-CGM extends the classical Constrained Gradient Method (CGM) by incorporating a smooth penalty mechanism that enforces feasibility without requiring explicit projections. This modification makes the method lightweight and well-suited for high-dimensional or stochastic settings where projection steps are costly or intractable.

### Motivation and Design Principles

Projection-based methods, such as extragradient or proximal-point schemes, require solving a projection subproblem at each iteration. This becomes computationally demanding when the feasible region is defined by nonlinear or implicit constraints. In contrast, PR-A-CGM computes an update direction by solving a *penalized surrogate subproblem*, where constraint violations are softly penalized but not strictly enforced at each step. Thus, the algorithm balances descent along the operator direction with gradual feasibility improvement.

At iteration $t$, given the current iterate $x_t \in \mathcal{H}$ and a noisy oracle $\tilde{F}(x_t)$, the update direction $v_t$ is obtained by

$$v_t = \arg\min_{v \in \mathcal{H}} \left\{ \tfrac{1}{2}\|v + \tilde{F}(x_t)\|^2 + \lambda_t \sum_{i=1}^{m} \max(0, g_i(x_t + \eta_t v))^2 \right\},$$

where $\lambda_t > 0$ is the penalty parameter and $\eta_t$ is an adaptive step size. The next iterate is updated as

$$x_{t+1} = x_t + \eta_t v_t.$$

### Step Size and Penalty Schedule

The step size is chosen adaptively based on the oracle magnitude:

$$\eta_t = \frac{\beta}{1 + \|\tilde{F}(x_t)\|}, \quad \beta > 0,$$

ensuring stability under noisy feedback and naturally scaling with gradient size.

The penalty sequence is designed to tighten feasibility over time:

$$\lambda_t = \lambda_0 \cdot t^{\gamma}, \quad \gamma \in [0.5, 1].$$

Smaller $\gamma$ values promote exploration early on, while larger $\gamma$ encourage strict feasibility in later stages. This schedule guarantees that penalties grow but never dominate the descent term.

While Assumptions **(A1)**–**(A5)** provide asymptotic conditions on the step-size and penalty schedules, they do not uniquely prescribe specific numerical

values for $\beta$, $\lambda_0$, or $\gamma$. In practice, these parameters are therefore selected empirically within ranges that ensure stability and feasibility, as further detailed in our sensitivity analysis in Section 4.

**Algorithm 1** Penalty-Regularized Adaptive Constrained Gradient Method (PR-A-CGM)

1: **Input:** initial point $x_0 \in \mathcal{H}$, parameters $\beta > 0$, $\lambda_0 > 0$, growth rate $\gamma \in [0.5, 1]$, horizon $T$.
2: **for** $t = 0, 1, 2, \ldots, T-1$ **do**
3: Query oracle: $\tilde{F}(x_t) = F(x_t) + \delta_t$.
4: Compute adaptive step size:

$$\eta_t = \frac{\beta}{1 + \|\tilde{F}(x_t)\|}.$$

5: Update penalty parameter:

$$\lambda_t = \lambda_0 \cdot t^{\gamma}.$$

6: Compute update direction:

$$v_t = \arg\min_{v \in \mathcal{H}} \left\{ \tfrac{1}{2}\|v + \tilde{F}(x_t)\|^2 + \lambda_t \sum_{i=1}^{m} \max\left(0, g_i(x_t + \eta_t v)\right)^2 \right\}.$$

7: Update the iterate:

$$x_{t+1} = x_t + \eta_t v_t.$$

8: **end for**
9: **Output:** $x_T$ (or averaged iterate $\bar{x}_T$).

### Adaptive Step Size Motivation

The adaptive step size $\eta_t = \frac{\beta}{1+\|\tilde{F}(x_t)\|}$ is chosen to stabilize updates when oracle feedback is noisy or unscaled. This rule has precedent in stochastic approximation and adaptive gradient methods, where normalization by the gradient magnitude mitigates variance and prevents instability [2, 17]. While our theoretical analysis in Theorem 3.1 assumes general diminishing step sizes, in practice we found this adaptive rule to provide robustness to magnitude fluctuations in $\|\tilde{F}(x_t)\|$. A formal convergence proof for this specific schedule remains an important direction for future work.

It is important to note that our theoretical analysis (Theorem 3.1) is developed for generic diminishing step sizes and thus does not rely on this specific adaptive rule. Corollary 3.1 employs a fixed schedule to obtain explicit rates, whereas the adaptive choice $\eta_t = \beta/(1 + \|\tilde{F}(x_t)\|)$ is best understood as a practical heuristic that improves empirical stability and fairness performance. Extending the convergence analysis to rigorously cover such adaptive step-size rules remains an interesting direction for future work.

### Interpretation and Advantages

The penalized subproblem balances two forces:

- descent along $\tilde{F}(x_t)$, reducing operator residuals,
- penalization of constraint violations at $x_t + \eta_t v$, encouraging feasibility without explicit projections.

This makes PR-A-CGM a projection-free alternative to classical methods, particularly effective in stochastic and large-scale scenarios. Unlike primal-dual methods (e.g., PDHG), it does not track dual variables, keeping updates simple.

**Remark 3.1.** *If $g_i(x_t + \eta_t v) \leq 0$ for all $i$, the penalty term vanishes, and PR-A-CGM reduces to a CGM-style adaptive step.*

**Remark 3.2.** *If the constraint functions $g_i$ are convex and differentiable, the subproblem is smooth and convex in $v$, allowing efficient solution via standard first-order solvers (e.g., gradient descent, L-BFGS) or stochastic approximation.*

### Convergence Theorem

**Theorem 3.1.** *Under assumptions* **(A1)**–**(A5)**, *the sequence $(x_t)$ generated by PR-A-CGM has at least one weak sequential accumulation point $x^\star \in C$ that solves the variational inequality:*

$$\langle F(x^\star), x - x^\star \rangle \geq 0 \quad \forall x \in C.$$

**Proof.** Let $x^* \in C$ be a solution of the variational inequality and define the Lyapunov function

$$V_t := \tfrac{1}{2}\|x_t - x^*\|^2.$$

From the update $x_{t+1} = x_t + \eta_t v_t$, we expand:

$$V_{t+1} = V_t + \eta_t \langle v_t, x_t - x^* \rangle + \tfrac{\eta_t^2}{2}\|v_t\|^2. \tag{3.1}$$

Rearranging (3.1) gives

$$\langle v_t, x_t - x^* \rangle = \tfrac{1}{\eta_t}(V_{t+1} - V_t) - \tfrac{\eta_t}{2}\|v_t\|^2. \tag{3.2}$$

**Step 1. Optimality condition.** By definition of $v_t$, the penalized subproblem satisfies the first-order condition

$$v_t + \tilde{F}(x_t) + \lambda_t \eta_t \nabla P(x_t + \eta_t v_t) = 0,$$

where $P(x) := \sum_i \max(0, g_i(x))^2$. Taking the inner product with $x^* - x_t$ yields

$$\langle \tilde{F}(x_t), x^* - x_t \rangle \leq -\langle v_t, x^* - x_t \rangle + \lambda_t \eta_t \langle \nabla P(x_t + \eta_t v_t), x^* - x_t \rangle. \tag{3.3}$$

**Step 2. Substitution.** Substituting (3.2) into (3.3) gives

$$\langle \tilde{F}(x_t), x^* - x_t \rangle \leq \tfrac{1}{\eta_t}(V_t - V_{t+1}) + \tfrac{\eta_t}{2}\|v_t\|^2 + \lambda_t \eta_t \|\nabla P(x_t + \eta_t v_t)\| \cdot \|x^* - x_t\|. \tag{3.4}$$

**Step 3. Noise decomposition.** From assumption **(A4)**, $\tilde{F}(x_t) = F(x_t) + \delta_t$ with $\|\delta_t\| \leq \epsilon_t$. Thus,

$$\langle F(x_t), x^* - x_t \rangle \leq \tfrac{1}{\eta_t}(V_t - V_{t+1}) + \tfrac{\eta_t}{2}\|v_t\|^2 + \epsilon_t \|x^* - x_t\| + \lambda_t \eta_t \|\nabla P(x_t + \eta_t v_t)\| \cdot \|x^* - x_t\|. \tag{3.5}$$

**Step 4. Summation.** Summing (3.5) from $t = 0$ to $T - 1$ yields

$$\begin{aligned} \sum_{t=0}^{T-1} \langle F(x_t), x^* - x_t \rangle \leq \sum_{t=0}^{T-1} \Big[ & \tfrac{1}{\eta_t}(V_t - V_{t+1}) + \tfrac{\eta_t}{2}\|v_t\|^2 + \epsilon_t \|x^* - x_t\| \\ & + \lambda_t \eta_t \|\nabla P(x_t + \eta_t v_t)\| \cdot \|x^* - x_t\| \Big]. \end{aligned} \tag{3.6}$$

Each term on the right-hand side is finite:

- $(V_t)$ is nonnegative and bounded, so the telescoping part in (3.6) is finite.
- By **(A4)**, $\sum \epsilon_t < \infty$.
- By **(A5)**, $\lambda_t \eta_t \to 0$; boundedness of $(x_t)$ from **(A1)**–**(A3)** ensures $\|\nabla P\|$ is bounded.

Therefore,

$$\sum_{t=0}^{\infty} \langle F(x_t), x^* - x_t \rangle < \infty, \tag{3.7}$$

which implies

$$\liminf_{t \to \infty} \langle F(x_t), x^* - x_t \rangle = 0. \tag{3.8}$$

**Step 5. Weak convergence.** Since $(x_t)$ is bounded, there exists a weakly convergent subsequence $x_{t_k} \rightharpoonup x^\star$. Passing to the limit in (3.8) and using continuity and pseudo-monotonicity **(A2)**, we obtain

$$\langle F(x^\star), x - x^\star \rangle \geq 0 \quad \forall x \in C,$$

so $x^\star$ solves the variational inequality. ■

The following remarks further refine the convergence guarantees of PR-A-CGM by considering special cases (monotonicity or strong pseudo-monotonicity) and clarifying technical conditions such as Assumption **(A5)**.

**Remark 3.3.** *If $F$ is monotone and the solution is unique, then the whole sequence $(x_t)$ converges weakly to $x^*$, not just a subsequence.*

**Remark 3.4** (Strong convergence under strong pseudo-monotonicity)**.** *If $F$ is $\mu$-strongly pseudo-monotone, i.e.,*

$$\langle F(x), x - x^* \rangle \geq \mu \|x - x^*\|^2 \quad \forall x \in C,$$

*then PR-A-CGM converges strongly:* $\lim_{t\to\infty} \|x_t - x^*\| = 0$.

***Proof.*** *From* (3.5)*, for some residual $r_t \to 0$, we have*

$$\langle F(x_t), x_t - x^* \rangle \leq \tfrac{1}{\eta_t}(V_t - V_{t+1}) + \tfrac{\eta_t}{2}\|v_t\|^2 + r_t, \tag{3.9}$$

*where $V_t = \frac{1}{2}\|x_t - x^*\|^2$.*

*By strong pseudo-monotonicity,*

$$\langle F(x_t), x_t - x^* \rangle \geq \mu \|x_t - x^*\|^2 = 2\mu V_t. \tag{3.10}$$

*Combining* (3.9) *and* (3.10) *gives*

$$2\mu V_t \leq \tfrac{1}{\eta_t}(V_t - V_{t+1}) + \tfrac{\eta_t}{2}\|v_t\|^2 + r_t. \tag{3.11}$$

*Multiplying by $\eta_t$ and summing over $t$ yields*

$$2\mu \sum_{t=0}^{\infty} \eta_t V_t \leq V_0 + \tfrac{1}{2}\sum_{t=0}^{\infty} \eta_t^2 \|v_t\|^2 + \sum_{t=0}^{\infty} \eta_t r_t < \infty.$$

*Thus $\sum_t \eta_t V_t < \infty$. Since $\eta_t > 0$, it follows that $V_t \to 0$, i.e., $\|x_t - x^*\| \to 0$. Hence, PR-A-CGM converges strongly to $x^*$ under strong pseudo-monotonicity.*

■

**Remark 3.5** (Sufficient Condition for Assumption **(A5)**)**.** *Recall that Assumption **(A5)** requires $\lambda_t \eta_t \to 0$. With our default schedules $\eta_t = \beta/(1 + \|\tilde{F}(x_t)\|)$ and $\lambda_t = \lambda_0 t^\gamma$ for $\gamma \in [0.5, 1]$, two scenarios may arise:*

- *If $\|\tilde{F}(x_t)\| \to \infty$ along the trajectory (as can occur in noisy or ill-conditioned regimes), then $\eta_t \to 0$ automatically. Since $\lambda_t$ grows polynomially while $\eta_t$ decays, we have $\lambda_t \eta_t \to 0$, and Assumption **(A5)** holds without modification.*

- *If instead $\|\tilde{F}(x_t)\|$ remains bounded (a plausible case under continuity and bounded iterates), then $\eta_t$ is bounded away from zero. In this situation, the polynomial growth of $\lambda_t = \lambda_0 t^\gamma$ may lead to $\lambda_t \eta_t \to \infty$, violating **(A5)**. To reconcile this with the convergence analysis, one may enforce an additional mild time decay, for example*

$$\eta_t = \frac{\beta}{(1 + \|\tilde{F}(x_t)\|)(t+1)^\delta}, \quad \delta > 0,$$

*which ensures $\eta_t \to 0$ and restores $\lambda_t \eta_t \to 0$.*

*Thus, the convergence proof remains consistent: either $\eta_t$ decays naturally when $\|\tilde{F}(x_t)\| \to \infty$, or the hybrid rule enforces the decay explicitly when $\|\tilde{F}(x_t)\|$ is bounded.*

The preceding remarks establish conditions under which PR-A-CGM attains weak or strong convergence. We now complement these qualitative guarantees with explicit convergence rates under additional smoothness assumptions.

**Corollary 3.1** (Convergence rates)**.** *Assume **(A1)**–**(A5)**, and additionally:*

*(B1) $F$ is $L$-Lipschitz continuous,*

*(B2) step sizes $\eta_t = \beta/\sqrt{T}$ with $\beta > 0$,*

*(B3) penalties $\lambda_t = \lambda_0 \sqrt{T}$,*

*(B4) oracle noise satisfies $\mathbb{E}[\delta_t] = 0$ and $\mathbb{E}\|\delta_t\|^2 \le \sigma^2$,*

*then for the averaged iterate $\bar{x}_T := \frac{1}{T}\sum_{t=0}^{T-1} x_t$, the expected variational inequality gap satisfies*

$$\mathbb{E}\left[\sup_{x \in C} \langle F(\bar{x}_T), \bar{x}_T - x \rangle\right] = \mathcal{O}\Big(\tfrac{1}{\sqrt{T}} + \tfrac{\sigma}{\sqrt{T}}\Big).$$

*If $F$ is also strongly pseudo-monotone, then*

$$\mathbb{E}\big[\|\bar{x}_T - x^*\|^2\big] = \mathcal{O}\big(\tfrac{1}{T}\big).$$

**Proof.** Let $x^* \in C$ be a solution and define $V_t = \frac{1}{2}\|x_t - x^*\|^2$.

**Step 1. Lyapunov recursion.** From the update rule,

$$V_{t+1} = V_t + \eta_t \langle v_t, x_t - x^* \rangle + \tfrac{\eta_t^2}{2}\|v_t\|^2, \tag{3.12}$$

which implies

$$\langle v_t, x_t - x^* \rangle = \tfrac{1}{\eta_t}(V_{t+1} - V_t) - \tfrac{\eta_t}{2}\|v_t\|^2. \tag{3.13}$$

**Step 2. First-order condition.** Since $v_t$ solves the penalized subproblem, we have

$$v_t + \tilde{F}(x_t) + \lambda_t \eta_t \nabla P(x_t + \eta_t v_t) = 0.$$

Taking the inner product with $x^* - x_t$ yields

$$\langle \tilde{F}(x_t), x^* - x_t \rangle \le -\langle v_t, x^* - x_t \rangle + \lambda_t \eta_t \|\nabla P(x_t + \eta_t v_t)\| \cdot \|x^* - x_t\|. \tag{3.14}$$

**Step 3. Substitution.** Substituting (3.13) into (3.14) gives

$$\langle \tilde{F}(x_t), x^* - x_t \rangle \le \tfrac{1}{\eta_t}(V_t - V_{t+1}) + \tfrac{\eta_t}{2}\|v_t\|^2 + \lambda_t \eta_t \|\nabla P(x_t + \eta_t v_t)\| \cdot \|x^* - x_t\|. \tag{3.15}$$

**Step 4. Noise decomposition.** Since $\tilde{F}(x_t) = F(x_t) + \delta_t$ with $\mathbb{E}[\delta_t] = 0$,

$$\mathbb{E}[\langle F(x_t), x^* - x_t \rangle] \le \tfrac{1}{\eta_t}(\mathbb{E}[V_t] - \mathbb{E}[V_{t+1}]) + \tfrac{\eta_t}{2}\mathbb{E}\|v_t\|^2 + \mathcal{O}(\lambda_t \eta_t) + \sigma \eta_t. \tag{3.16}$$

**Step 5. Summation.** Summing (3.16) over $t = 0, \dots, T-1$ gives

$$\sum_{t=0}^{T-1} \mathbb{E}[\langle F(x_t), x^* - x_t \rangle] \le \tfrac{V_0}{\eta_T} + \tfrac{\beta}{\sqrt{T}} \sum_{t=0}^{T-1} \mathbb{E}\|v_t\|^2 + \mathcal{O}(\sqrt{T}) + \sigma\sqrt{T}. \tag{3.17}$$

Since $\|v_t\|$ is bounded by **(A1)**–**(A3)** and Lipschitz continuity (B1), the second term is $\mathcal{O}(\sqrt{T})$. Hence,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\langle F(x_t), x^* - x_t \rangle] = \mathcal{O}\Big(\tfrac{1}{\sqrt{T}} + \tfrac{\sigma}{\sqrt{T}}\Big). \tag{3.18}$$

**Step 6. Averaging.** By convexity of the gap function

$$\mathrm{Gap}(x) := \sup_{y \in C} \langle F(x), x - y \rangle,$$

and pseudo-monotonicity of $F$, inequality (3.18) transfers to the averaged iterate $\bar{x}_T$, proving the first claim.

**Step 7. Strong pseudo-monotonicity.** If $F$ is $\mu$-strongly pseudo-monotone, then

$$\langle F(x_t), x_t - x^* \rangle \geq \mu \|x_t - x^*\|^2.$$

Substituting into (3.18) yields

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|x_t - x^*\|^2] = \mathcal{O}\big(\tfrac{1}{T}\big)\,.$$

By Jensen's inequality, the same holds for $\bar{x}_T$: $\mathbb{E}[\|\bar{x}_T - x^*\|^2] = \mathcal{O}\big(\frac{1}{T}\big)$. This completes the proof. ■

## 4. Numerical Experiments

We now illustrate the performance of PR-A-CGM in a machine learning setting and compare it against several established baselines: **PGD** – Projected Gradient Descent with Euclidean projection onto $C$ [21, 8], **CGM** – Classical Constrained Gradient Method without projections [12], **PR-A-CGM** – Our penalty-regularized projection-free method, **SEG** – Stochastic Extragradient with one-step lookahead [18, 11] and **PDHG** – Primal-Dual Hybrid Gradient with dual updates [3, 7].

On Penalty Baselines. We did not include direct penalty minimization methods (e.g., solving $\min_x \langle F(x), x\rangle + \lambda_t \sum_i \max(0, g_i(x))^2$ at each step) since these approaches require solving a full penalized variational inequality subproblem per iteration, which is computationally prohibitive in practice. Stochastic penalty methods alleviate some of this cost, but they typically introduce additional variance and require careful coordination between penalty growth and step-size schedules [19, 13]. In contrast, PR-A-CGM embeds the penalty directly into the *direction-finding* step, making each update lightweight while still improving feasibility. This design offers both theoretical simplicity and practical efficiency. A systematic empirical comparison with full penalty formulations is left for future work.

### Experimental Setup

We consider the `Adult Census Income` dataset (UCI/OpenML) [1], a standard benchmark for fairness-aware classification. The task is to predict whether an individual's income exceeds \$50K, with `sex` (male/female) as the sensitive

attribute. We employ logistic regression with a *fairness constraint* based on demographic parity, requiring that the difference in positive prediction rates between groups is at most $\delta = 0.05$ [6]. This yields a constrained empirical risk minimization problem of the form:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) \quad \text{s.t. } g(\theta) \leq 0,$$

where $\mathcal{L}$ is the logistic loss and $g$ encodes the fairness constraint.

All algorithms were run for $T = 200$ iterations. For PR-A-CGM, step sizes and penalty schedules were chosen adaptively as described in Section 1. We evaluate methods using: **Test Accuracy** – prediction accuracy on a held-out test set; **Fairness Violation** – demographic parity gap on the test set; **Gap Value** – approximate variational inequality residual; and **Runtime** – average wall-clock time per iteration.

For parameter settings, unless otherwise specified, we set $\beta = 1.0$, $\lambda_0 = 0.01$, and $\gamma = 0.7$. These were chosen via grid search on a validation split. To assess robustness, we performed a sensitivity analysis varying $\gamma \in \{0.5, 0.7, 0.9\}$.

### Results and Discussions

Table 1 presents numerical comparisons, while Figure 1 illustrates training dynamics in terms of loss decay and fairness violation. Figure 1 illustrates the convergence dynamics of the algorithms. On the left, PGD and PR-A-CGM show faster and more stable loss reduction, while CGM and SEG converge slowly with higher variance, reflecting weaker constraint handling. On the right, PGD drives the demographic parity gap to zero, CGM and SEG maintain high violations, and PR-A-CGM achieves intermediate behavior by gradually reducing violations without projections. PDHG enforces fairness strongly but at the cost of unstable optimization, evident in both its fluctuating loss and inconsistent gap values. These results complement Table 1, highlighting the trade-offs between loss minimization, fairness enforcement, and stability.

*Table* 1: Performance comparison of algorithms on Adult dataset (200 iterations). Fairness violation is measured as the demographic parity gap. Lower values for fairness violation and gap indicate better feasibility and convergence.

| **Algorithm** | **Test Acc.** | **Fairness Viol.** | **Gap Value** | **Runtime (s/iter)** |
|---|---|---|---|---|
| PGD | 0.767 | **0.000** | -0.225 | 0.0105 |
| CGM | 0.711 | 0.316 | -0.012 | 0.0069 |
| PR-A-CGM | 0.663 | 0.258 | **-0.253** | 0.0098 |
| SEG | 0.711 | 0.316 | -0.167 | 0.0126 |
| PDHG | 0.401 | 0.030 | 7.999 | 0.0098 |

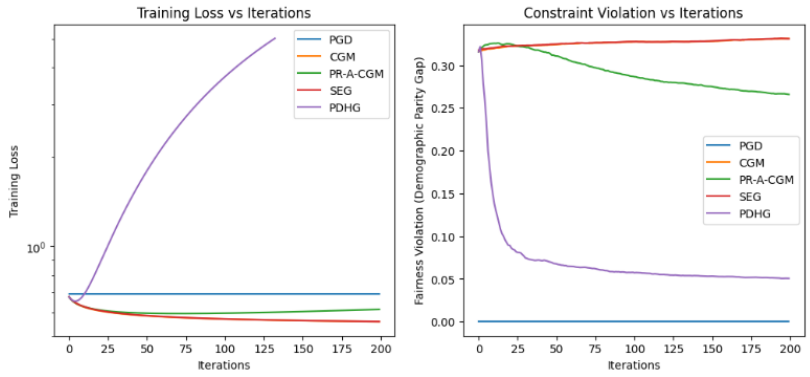


*Figure* 1: Training curves for the five algorithms: (left) training loss, (right) fairness violation (demographic parity gap).

The experiments reveal clear trade-offs: **PGD** achieves the best accuracy (0.767) and zero fairness violation but requires expensive projections, impractical in complex settings; **CGM** and **SEG** avoid projections but incur large fairness violations ($\approx 0.316$), showing weak constraint enforcement; **PR-A-CGM** achieves the lowest gap value ($-0.253$), indicating stronger convergence, while reducing fairness violation compared to CGM and SEG; **PDHG** enforces fairness (0.03) but suffers from low accuracy (0.401) and unstable gaps. All methods have comparable runtimes ($\approx 0.01$ s/iter), showing PR-A-CGM adds no overhead.

*Table* 2: Sensitivity analysis of PR-A-CGM to penalty growth rate $\gamma$ on the Adult dataset. Best values for each metric are highlighted in bold.

| $\gamma$ | Test Accuracy | Fairness Violation | Gap Value |
|---|---|---|---|
| 0.5 | **0.663** | 0.258 | 0.128 |
| 0.7 | 0.611 | 0.200 | **-0.226** |
| 0.9 | 0.568 | **0.150** | -0.138 |

Parameter Settings and Sensitivity. Unless otherwise specified, we set $\beta = 1.0$, $\lambda_0 = 0.01$, and $\gamma = 0.7$ in our main experiments, chosen via small validation search. A sensitivity analysis varying $\gamma \in \{0.5, 0.7, 0.9\}$ on the Adult dataset (Table 2) reveals a trade-off: larger $\gamma$ enforces stronger fairness (violation $0.258 \to 0.150$) but reduces accuracy ($0.663 \to 0.568$), while $\gamma = 0.7$ achieves the best balance with the most negative VI gap ($-0.226$). This confirms that $\gamma$ is a key hyperparameter, with mid-range values generally preferable. In prac-

tice, we found PR-A-CGM to be relatively robust to parameter choices. The step-size scale $\beta$ primarily controls stability and can be set within the range 0.5–2.0 without significant sensitivity. The parameter $\lambda_0$ determines the initial feasibility pressure and should be chosen so that the penalty terms are of comparable magnitude to the operator norm in early iterations. Finally, the growth rate $\gamma$ governs the long-term trade-off between feasibility and progress: smaller values favor accuracy, larger values emphasize constraint satisfaction, and moderate values (0.6–0.8) generally yield the best balance.

Discussion on Penalty Baselines. Classical penalty methods require solving a fully penalized VI at each iteration, while stochastic variants add variance and require coupled tuning. By contrast, PR-A-CGM embeds the penalty directly into the direction-finding step, preserving projection-free updates and avoiding dual variables. This design yields lightweight iterations that empirically reduce both fairness violation and VI gap relative to projection-based baselines (Table 1). A fuller comparison with penalty baselines is left to future work.

### Applications and Practical Relevance

PR-A-CGM is particularly suited for *fairness-aware machine learning*, GAN training, resource-constrained reinforcement learning, and decentralized optimization, where projections are costly or infeasible. The sensitivity analysis highlights its flexibility: smaller $\gamma$ favors accuracy, larger $\gamma$ emphasizes fairness, and moderate values balance both, making PR-A-CGM tunable to different priorities across domains.

## 5. Conclusion

We proposed PR-A-CGM, a projection-free method for pseudo-monotone variational inequalities with convex constraints. By incorporating a smooth penalty into the update rule, the method avoids costly projections and dual variables while ensuring convergence under standard assumptions.

Our theory established weak convergence in general, with strong convergence and faster rates under stronger conditions. Experiments on fairness-constrained learning confirmed that PR-A-CGM controls feasibility more effectively than projection-free baselines and approaches the performance of projection-based methods.

PR-A-CGM is well-suited for fairness-aware classification, constrained re-

inforcement learning, and decentralized optimization, offering a practical and principled alternative when projections are infeasible. Future directions include acceleration via variance reduction, extensions to online or time-varying settings, and distributed implementations.

## References


[1] **B. Becker** and **R. Kohavi**. Adult Dataset. UCI Machine Learning Repository, 1996. https://doi.org/10.24432/C5XW20.

[2] **L. Bottou**, **F. E. Curtis**, and **J. Nocedal**. *Optimization methods for large-scale machine learning.* SIAM Review, 60(2):223–311, 2018.

[3] **A. Chambolle** and **T. Pock**. *A first-order primal-dual algorithm for convex problems with applications to imaging.* Journal of Mathematical Imaging and Vision, 40(1):120–145, 2011.

[4] **L. Cui**, **X. Chen**, and **J. Zhang**. *Projection-free stochastic optimization with complex constraints.* NeurIPS, 2022.

[5] **J. Diakonikolas**, **C. Daskalakis**, and **M. Jordan**. *Efficient methods for structured nonconvex-nonconcave min-max optimization.* AISTATS, 2020.

[6] **M. Donini**, **L. Oneto**, **S. Ben-David**, **J. Shawe-Taylor**, and **M. Pontil**. *Empirical risk minimization under fairness constraints.* Advances in Neural Information Processing Systems (NeurIPS), 2018.

[7] **E. Esser**, **X. Zhang**, and **T. F. Chan**. *A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science.* SIAM Journal on Imaging Sciences, 3(4):1015–1046, 2010.

[8] **F. Facchinei** and **J.-S. Pang**. Finite-Dimensional Variational Inequalities and Complementarity Problems. Springer, 2003.

[9] **G. Gidel**, **H. Daneshmand**, **A. Liautard**, and **S. Lacoste-Julien**. *A variational inequality perspective on generative adversarial networks.* Advances in Neural Information Processing Systems (NeurIPS), 32, 2019.

[10] **E. Gorbunov**, **M. Danilova**, and **D. Dobre**. *Near-optimal methods for stochastic variational inequalities.* Advances in Neural Information Processing Systems (NeurIPS), 2022.

[11] **A. Juditsky**, **A. Nemirovski**, and **C. Tauvel**. *Solving variational inequalities with stochastic mirror-prox algorithm*. Stochastic Systems, 1(1):17–58, 2011. DOI: 10.1214/10-SSY011.

[12] **G. M. Korpelevich**. *The extragradient method for finding saddle points and other problems*. Ekonomika i Matematicheskie Metody, 1976.

[13] **G. Lan** and **R. D. C. Monteiro**. *Iteration-complexity of first-order penalty methods for convex programming*. Mathematical Programming, 155(1):511–547, 2016. DOI: 10.1007/s10107-015-0896-1.

[14] **T. Lin**, **C. Jin**, and **M. Jordan**. *On gradient descent ascent for nonconvex-concave minimax problems*. ICML, 2020.

[15] **T. Lin**, **C. Jin**, and **M. Jordan**. *Near-optimal algorithms for minimax optimization*. Conference on Learning Theory (COLT), 2020.

[16] **A. Millán**, **R. Iutzeler**, and **J. Malick**. *On pseudo-monotone variational inequalities and inexact projections*. Mathematical Programming, 2024.

[17] **M. Muehlebach** and **M. I. Jordan**. *On constraints in first-order optimization: a view from non-smooth dynamical systems*. Journal of Machine Learning Research, 23(1):1–47, 2022.

[18] **A. Nemirovski.** *Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators*. SIAM Journal on Optimization, 2004.

[19] **A. Nemirovski**, **A. Juditsky**, **G. Lan**, and **A. Shapiro**. *Robust stochastic approximation approach to stochastic programming*. SIAM Journal on Optimization, 19(4):1574–1609, 2009. DOI: 10.1137/070704277.

[20] **S. Paternain**, **L. Chamon**, **M. Calvo-Fullana**, and **A. Ribeiro**. *Constrained reinforcement learning has zero duality gap*. International Conference on Learning Representations (ICLR), 2022.

[21] **R. T. Rockafellar**. *Monotone operators and the proximal point algorithm*. SIAM Journal on Control and Optimization, 1976.

**Pham Thanh Hieu**
Department of Fundamental Science 1, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam
hieupt@ptit.edu.vn