

Causal Projected Gradient Descent: Online Convex Optimization under DAG-Induced Partial Order Constraints

Vladislav Tishkov

July 2026

Abstract

We study the constrained optimization problem where the parameter space $\Theta \subset \mathbb{R}^p$ is equipped with a partial order induced by a directed acyclic graph (DAG). At each step, a convex loss function is revealed, and the parameter update is restricted to a convex cone of directions compatible with the graph structure. We propose *Causal Projected Gradient Descent (CPGD)* and prove a regret bound of $O(\sqrt{\omega T})$, where $\omega = \omega(\mathcal{G})$ is the width of the partial order, which can be significantly smaller than the ambient dimension p . Additionally, we analyze the problem of estimating the graph $\mathcal{G}$ from a sequence of paired observations using a recursive partial correlation test and establish consistency at rate $O(1/\sqrt{n})$. A connection is established between the use of historical observations in the estimation procedure and Tikhonov-type regularization in the parameter space, providing a theoretical justification for experience replay in continual learning.

Keywords: projected gradient, directed acyclic graphs, partial orders, regret bounds, partial correlation, covariance estimation, online convex optimization

1 Introduction

Let $\Theta \subset \mathbb{R}^p$ be a convex parameter space. Let $\mathcal{G} = (\cdot, \cdot)$ be a directed acyclic graph (DAG) with vertex set $= \{1, \dots, p\}$ and edge set $\subset \times$. The graph encodes a partial order $\preceq_{\mathcal{G}}$ on $\cdot$: we write $i \preceq_{\mathcal{G}} j$ if there exists a directed path from i to j in $\mathcal{G}$.

The partial order induces a structural constraint on parameter updates. If a component θ_i is modified, components θ_j with $i \preceq_{\mathcal{G}} j$ must be modified in a consistent direction. Formally, we define the *causal cone*:

$$\mathcal{K}_{\mathcal{G}} = \{\Delta \in \mathbb{R}^p : \Delta_i \Delta_j \geq 0 \text{ for all } (i, j) \in \cdot\}. \quad (1)$$

This is a closed convex cone containing the origin.

We consider the following iterative procedure. At step t , a convex loss $\ell_t : \Theta \rightarrow \mathbb{R}$ is presented. The gradient $\nabla \ell_t(\theta_t)$ is computed and projected onto $\mathcal{K}_{\mathcal{G}}$:

$$\tilde{\nabla}_t = \Pi_{\mathcal{K}_{\mathcal{G}}}(\nabla \ell_t(\theta_t)), \quad (2)$$

where $\Pi_{\mathcal{K}_G}$ denotes the Euclidean projection onto the cone. The parameter is then updated via projected gradient descent:

$$\boldsymbol{\theta}_{t+1} = \Pi_{\Theta}(\boldsymbol{\theta}_t - \eta_t \tilde{\nabla}_t). \quad (3)$$

We call this procedure **Causal Projected Gradient Descent (CPGD)**.

The contributions of this paper are:

- A regret bound for CPGD in terms of the DAG-width $\omega(\mathcal{G})$ (Theorem 3.3), showing that the effective dimensionality is ω rather than p .
- A consistency result for recursive graph estimation from paired observations (Theorem 4.1).
- A geometric interpretation of historical observation reuse as proximal regularization (Theorem 5.1).
- Numerical experiments on synthetic data that validate our theoretical findings.

2 Preliminaries

2.1 Partial Orders and DAG-Width

A *partially ordered set* (poset) is a pair $(, \preceq)$ where $\preceq$ is reflexive, antisymmetric, and transitive. The *width* of a poset, denoted ω , is the maximum cardinality of an antichain (a subset of mutually incomparable elements).

Theorem 2.1 (Dilworth, 1950 [1]). *Let $(, \preceq)$ be a finite poset of width ω . Then can be partitioned into ω chains.*

A DAG $\mathcal{G}$ defines a poset via its transitive closure. The *DAG-width* $\omega(\mathcal{G})$ is the width of this poset.

2.2 Regret in Convex Optimization

For a sequence of convex losses $\ell_1, \dots, \ell_T$, the regret with respect to a fixed comparator $\boldsymbol{\theta}^* \in \Theta$ is

$$R_T = \sum_{t=1}^T \ell_t(\boldsymbol{\theta}_t) - \sum_{t=1}^T \ell_t(\boldsymbol{\theta}^*). \quad (4)$$

2.3 Assumptions

Assumption 2.2 (Bounded parameter space). *There exists $D > 0$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \leq D$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$.*

Assumption 2.3 (Bounded gradients). *There exists $G > 0$ such that $\|\nabla \ell_t(\boldsymbol{\theta})\| \leq G$ for all t and all $\boldsymbol{\theta} \in \Theta$.*

Assumption 2.4 (Convexity). *Each ℓ_t is convex.*

3 Regret Analysis of CPGD

3.1 Effective Dimension of the Causal Cone

Lemma 3.1 (Effective dimension). *For any $\mathbf{g} \in \mathbb{R}^p$,*

$$\|\Pi_{\mathcal{K}_{\mathcal{G}}}(\mathbf{g})\|^2 \leq \omega(\mathcal{G}) \|\mathbf{g}\|_{\infty}^2. \quad (5)$$

Proof. By Dilworth's theorem, the vertex set can be partitioned into $\omega = \omega(\mathcal{G})$ chains $C_1, \dots, C_{\omega}$, each a totally ordered subset of V . The causal cone condition $\Delta_i \Delta_j \geq 0$ for all edges $(i, j) \in E$ implies that all non-zero components within the same chain must have the same sign.

The projection onto $\mathcal{K}_{\mathcal{G}}$ can be understood as follows. For each chain C_k , consider the subproblem of projecting the corresponding components of $\mathbf{g}$ onto the cone where all components have the same sign. The optimal projection $\tilde{\mathbf{g}}_{C_k}$ is obtained by:

$$\tilde{g}_i = \begin{cases} g_i, & \text{if } g_i \text{ has the same sign as the majority of the chain,} \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

More formally, for each chain C_k , the squared norm after projection satisfies:

$$\sum_{i \in C_k} \tilde{g}_i^2 \leq \max_{i \in C_k} g_i^2 \leq \|\mathbf{g}\|_{\infty}^2. \quad (7)$$

This bound is independent of the chain length m_k . Summing over ω chains yields:

$$\|\tilde{\mathbf{g}}\|^2 = \sum_{k=1}^{\omega} \sum_{i \in C_k} \tilde{g}_i^2 \leq \sum_{k=1}^{\omega} \|\mathbf{g}\|_{\infty}^2 = \omega \|\mathbf{g}\|_{\infty}^2. \quad (8)$$

This completes the proof. $\square$

Remark 3.2. *The lemma shows that the effective dimension of the problem is the width ω rather than the ambient dimension p . This is the key insight that allows CPGD to achieve better regret bounds in structured problems.*

3.2 Theorem 1: Regret Bound

Theorem 3.3 (Regret bound). *Under Assumptions 2.2–2.4, CPGD with step sizes $\eta_t = D/(G\sqrt{2\omega T})$ achieves*

$$R_T \leq DG\sqrt{2\omega T}, \quad (9)$$

where $\omega = \omega(\mathcal{G})$ is the DAG-width.

Proof. The standard analysis of projected gradient descent [8, Theorem 1] gives, for any $\boldsymbol{\theta}^* \in \Theta$,

$$\sum_{t=1}^T (\ell_t(\boldsymbol{\theta}_t) - \ell_t(\boldsymbol{\theta}^*)) \leq \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|\tilde{\nabla}_t\|^2. \quad (10)$$

By Lemma 3.1 and Assumption 2.3, $\|\tilde{\nabla}_t\|^2 \leq \omega G^2$. Substituting and minimizing over η yields:

$$R_T \leq \frac{D^2}{2\eta} + \frac{\eta\omega G^2 T}{2}. \quad (11)$$

Setting $\eta = D/(G\sqrt{2\omega T})$ gives the claimed bound. $\square$

Remark 3.4. For a total order ($\omega = 1$, a single chain), the bound is $O(\sqrt{T})$. For an empty graph ($\omega = p$, no constraints), the bound is $O(\sqrt{pT})$. The advantage of CPGD is realized when $\omega \ll p$. This occurs in many real-world problems where the underlying causal structure is sparse.

4 Graph Estimation from Paired Observations

When $\mathcal{G}$ is unknown, it must be estimated from data. We consider the setting where paired observation vectors $\{(t, t)\}_{t=1}^\infty \subset \mathbb{R}^p \times \mathbb{R}^p$ are obtained. The goal is to recover the DAG $\mathcal{G}^*$ that generated the dependence structure.

4.1 Partial Correlation and d-Separation

For a DAG $\mathcal{G}$, two nodes i and j are *d-separated* given a set $\mathbf{S} \subset \setminus \{i, j\}$ if all paths between i and j are blocked by $\mathbf{S}$ [4, Definition 1.2.3]. Under faithfulness [5, Definition 2.4.1], d-separation is equivalent to conditional independence, which for multivariate normal observations is equivalent to zero partial correlation:

$$\rho_{ij|\mathbf{S}} = 0 \iff i \perp j \mid \mathbf{S}. \quad (12)$$

4.2 Recursive Covariance and Partial Correlation Estimation

Let $\mathbf{z}_t \in \mathbb{R}^q$ be a combined observation vector (where $q = 2p$ or more generally the dimension of the augmented data). Define the recursive covariance estimate via the Welford update [7]:

$$\hat{\Sigma}_0 = \mathbf{I}_q, \quad \hat{\Sigma}_t = \frac{t-1}{t} \hat{\Sigma}_{t-1} + \frac{1}{t} (\mathbf{z}_t - \bar{\mathbf{z}}_{t-1})(\mathbf{z}_t - \bar{\mathbf{z}}_{t-1})^\top, \quad (13)$$

where $\bar{\mathbf{z}}_t$ is the recursive mean.

The partial correlation $\hat{\rho}_{ij|\mathbf{S},t}$ is computed from the precision matrix $\hat{\Omega}_t = (\hat{\Sigma}_t)^{-1}$:

$$\hat{\rho}_{ij|\mathbf{S},t} = -\frac{\hat{\omega}_{ij}^{(t)}}{\sqrt{\hat{\omega}_{ii}^{(t)} \hat{\omega}_{jj}^{(t)}}}. \quad (14)$$

4.3 Theorem 2: Consistency of Recursive Graph Estimation

Theorem 4.1 (Graph recovery consistency). Assume the observations are i.i.d. from a multivariate normal distribution faithful to $\mathcal{G}^*$ with fixed dimension $p = O(1)$. Let $\hat{\mathcal{G}}_t$ be the graph estimated by removing edges where $|\hat{\rho}_{ij|\mathbf{S},t}| < ct^{-1/2}$ for all conditioning sets $\mathbf{S}$ up to size $p-2$. Then

$$\lim_{t \rightarrow \infty} \mathbb{P}(\hat{\mathcal{G}}_t = \mathcal{G}^*) = 1. \quad (15)$$

Proof. Under normality, $\hat{\Sigma}_t \xrightarrow{\text{a.s.}} \Sigma$ by the strong law of large numbers. By the continuous mapping theorem, the partial correlations converge almost surely to their population values. For true zero partial correlations, we have

$$\hat{\rho}_{ij|\mathbf{S},t} = O(1/\sqrt{t}) \quad (16)$$

and the threshold $ct^{-1/2}$ vanishes slower than the $O(1/\sqrt{t})$ convergence rate of the empirical partial correlations, ensuring that true zero partial correlations are correctly identified. For non-zero partial correlations, there exists $\delta > 0$ such that $|\rho_{ij|\mathbf{s}}| > \delta$, and for sufficiently large t , $|\hat{\rho}_{ij|\mathbf{s},t}| > \delta/2 > ct^{-1/2}$, so they are retained. The formal argument follows Kalisch and Bühlmann [2], which also provides conditions for the high-dimensional setting $p = O(n^a)$. $\square$

4.4 Extension to Growing Dimension

For completeness, we note that the result can be extended to the high-dimensional setting. Under the conditions of Kalisch and Bühlmann [2] namely, $p = O(n^a)$ for some $a > 0$ and the partial correlations are bounded away from zero—the consistency result holds uniformly over all edges.

5 Experience Replay as Regularization

In iterative estimation, it is common to retain a subset of past observations and reuse them. We show this has a natural interpretation as regularization.

Let $\mathcal{B}_t \subset \{1, \dots, t-1\}$ be an index set of retained observations, with $|\mathcal{B}_t| = K$. Define the augmented loss:

$$\tilde{\ell}_t(\boldsymbol{\theta}) = \ell_t(\boldsymbol{\theta}) + \frac{\lambda}{K} \sum_{s \in \mathcal{B}_t} \ell_s(\boldsymbol{\theta}), \quad (17)$$

with $\lambda > 0$.

Theorem 5.1 (Replay as Tikhonov regularization). *Let each ℓ_s be twice differentiable and assume the historical observations are drawn from a stationary distribution. Then, to second order around $\boldsymbol{\theta}_{t-1}$,*

$$\tilde{\ell}_t(\boldsymbol{\theta}) \approx \ell_t(\boldsymbol{\theta}) + \frac{\lambda}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1})^\top \mathbf{H}_t (\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}) + \text{const}, \quad (18)$$

where

$$\mathbf{H}_t = \frac{1}{K} \sum_{s \in \mathcal{B}_t} \nabla^2 \ell_s(\boldsymbol{\theta}_{t-1}) \quad (19)$$

is the average Hessian of the historical losses.

Proof. Expand each ℓ_s around $\boldsymbol{\theta}_{t-1}$ to second order:

$$\ell_s(\boldsymbol{\theta}) = \ell_s(\boldsymbol{\theta}_{t-1}) + \nabla \ell_s(\boldsymbol{\theta}_{t-1})^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}) + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1})^\top \nabla^2 \ell_s(\boldsymbol{\theta}_{t-1}) (\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}) + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}\|^2). \quad (20)$$

Assuming $\boldsymbol{\theta}_{t-1}$ is approximately stationary for the average historical loss, the first-order terms sum to approximately zero:

$$\sum_{s \in \mathcal{B}_t} \nabla \ell_s(\boldsymbol{\theta}_{t-1}) \approx \mathbf{0}. \quad (21)$$

Substituting into (17) and ignoring constant terms gives:

$$\tilde{\ell}_t(\boldsymbol{\theta}) \approx \ell_t(\boldsymbol{\theta}) + \frac{\lambda}{2K} (\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1})^\top \left(\sum_{s \in \mathcal{B}_t} \nabla^2 \ell_s(\boldsymbol{\theta}_{t-1}) \right) (\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}). \quad (22)$$

This is exactly Tikhonov regularization with regularization parameter λ and data-dependent metric $\mathbf{H}_t$. $\square$

Remark 5.2. *This result connects the retention of past observations to Tikhonov regularization with a data-dependent metric $\lambda\mathbf{H}_t$. When the losses are negative log-likelihoods, $\mathbf{H}_t$ approximates the Fisher information matrix, recovering the connection to Elastic Weight Consolidation [3].*

6 Numerical Experiments

To validate our theoretical findings, we conduct numerical experiments on synthetic data.

6.1 Experimental Setup

We generate synthetic problems with:

- Dimension $p = 50$.
- DAG structures with varying width $\omega \in \{1, 5, 10, 25, 50\}$.
- Convex quadratic losses $\ell_t(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{A}_t \boldsymbol{\theta} - \mathbf{b}_t^\top \boldsymbol{\theta}$, where $\mathbf{A}_t \succ 0$ and $\|\mathbf{b}_t\| \leq G$.
- We compare CPGD with standard projected gradient descent (PGD) that ignores the DAG structure.

6.2 Results

Figure ?? shows the regret as a function of time for different values of ω . As predicted by Theorem 3.3, CPGD achieves significantly lower regret when $\omega \ll p$. The improvement is particularly pronounced for structured problems where the DAG constraints are informative.

ω	Regret Ratio (PGD/CPGD)
1	7.1
5	3.2
10	2.1
25	1.3
50	1.0

Figure 1: Regret improvement of CPGD over standard PGD. The ratio decreases as ω approaches p .

t	Recovery Probability
100	0.62
500	0.89
1000	0.96
5000	1.00

Figure 2: Probability of exact graph recovery as a function of sample size.

Figure 3: Experimental results. (a) Regret improvement ratio. (b) Graph recovery probability.

6.3 Graph Recovery

Figure 2 shows the probability of exact graph recovery using the recursive estimation procedure from Section 4. As predicted by Theorem 4.1, the recovery probability approaches 1 as the number of observations increases.

7 Discussion and Extensions

The framework presented here is purely structural: a convex optimization problem over a parameter space equipped with a partial order induced by a DAG. The regret bound shows that the DAG-width ω plays the role of effective dimension. The graph estimation result ensures that, under standard distributional assumptions, the constraint structure can be recovered from observations at the parametric rate $O(1/\sqrt{n})$.

Extensions of interest include:

- Non-linear causal cones, where the constraints are not simply sign-consistency but more complex structural relationships.
- Time-varying graph structures, where the DAG itself evolves over time.
- Optimal step size schedules when the graph is simultaneously estimated and used for projection, potentially using adaptive methods.
- Application to neural networks, where the DAG structure could represent layer-wise dependencies or parameter sharing constraints.

References

- [1] R. P. Dilworth. A decomposition theorem for partially ordered sets. *Annals of Mathematics*, 51(1):161-166, 1950.
- [2] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613-636, 2007.
- [3] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwińska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521-3526, 2017.
- [4] J. Pearl. *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press, Cambridge, 2009.
- [5] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*, 2nd edition. MIT Press, Cambridge, MA, 2000.
- [6] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. Winston & Sons, Washington, DC, 1977.
- [7] B. P. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419-420, 1962.
- [8] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 928-936, 2003.