B2P-Corr: Batch-to-Population Gradient Estimators for Non-Decomposable Correlation Losses

Jingquan Yan, Yuwei Miao, Peiran Yu, Junzhou Huang*
Department of Computer Science and Engineering
University of Texas at Arlington
Arlington, TX, USA
{jingquan.yan,yuwei.miao,peiran.yu,jzhuang}@uta.edu

Abstract

Correlation metrics such as the Pearson Correlation Coefficient (PCC) and Concordance Correlation Coefficient (CCC) are standard regression metrics and learning objectives, but their non-decomposable nature poses a challenge for stochastic optimization. Since these metrics depend on dataset-level moments (mean, variance, covariance), naive mini-batch training with local statistics yields biased, high-variance gradients whose expectation does not align with the population gradient. To bridge this "batch-to-population" (B2P) gap, we propose **B2P-Corr**, a novel framework that transforms correlation objectives into decomposable pseudo-losses driven by lightweight global "moment sketches". We introduce two efficient variants: **B2P-EMA**, which utilizes exponential moving averages and stop-gradient operators to ensure asymptotic gradient tracking with $O(B)$ cost (B is batch size); and **B2P-U**, which uses sampled cross-batch U-statistic estimators and a reservoir-based stabilization with an $O(B)$ implementation. Theoretically, we quantify the $O(B^{-1})$ bias of naive mini-batch optimization, prove a moving-target tracking result for B2P-EMA under a fast-sketch / slow-parameter regime, and provide an idealized control-variate analysis for B2P-U. Experiments on two synthetic and three real-world datasets demonstrate B2P-Corr achieves stable optimization and higher correlation than naive mini-batch optimization with negligible overhead.

1 Introduction

Stochastic optimization (e.g., SGD [18]) are most effective for *decomposable* objectives, where the population risk is an expectation of per-example losses and mini-batch gradients are unbiased estimators of the population gradient. This assumption breaks for *non-decomposable* objectives whose values depend on dataset-level statistics or cross-example interactions. Well-known examples include AUC/F-measures in classification [19] and correlation-based regression objectives such as the Pearson correlation coefficient (PCC) [20] and concordance correlation coefficient (CCC) [9].

Correlation metrics are natural and crucial when relative trends matter more than absolute scale, e.g., in transcriptomics or human-annotated regression tasks where annotator-dependent scale shifts make absolute errors less informative [21]. However, they depend on global means, variances, and covariances. Replacing these population moments with mini-batch estimates makes the objective batch-dependent, so the resulting gradients are noisy and biased, especially in the small-batch regime.

Existing methods are incomplete. Larger batches reduce moment noise but are often memory-intensive and may hurt generalization [7, 8]. Beyond large-batch strategy, prior work on non-decomposable objectives includes primal–dual or constrained formulations [19], general stochastic compositional

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

optimization [24], coupled/compositional methods for AP and AUPRC-style objectives [23, 15], and DRO-based formulations for partial AUC [28]. These advances are important, but they mostly target classification metrics and do not directly resolve mini-batch correlation optimization in regression within a standard SGD pipeline.

We view this mismatch as a *batch-to-population* (B2P) gap: the target correlation objective is global, but training only observes noisy local statistics. Our key observation is that, although PCC/CCC are non-decomposable as scalar objectives, their population gradients admit decomposable per-example forms whose coefficients depend only on a few global moments.

Contributions. We propose **B2P-Corr** (Figure 1), which bridges this B2P gap by decoupling moment estimation from backpropagation via stop-gradient pseudo-losses and lightweight moment sketches. Specifically, we

(i) derive population gradients of PCC and CCC as decomposable per-example expectations depending on global prediction moments, and implement them via pseudo-losses with stop-gradient sketches.

(ii) introduce **B2P-EMA**, which tracks global moments with EMA, and under a standard fast-sketch / slow-parameter regime, it yields plug-in gradients that asymptotically track the moving population target at $O(B)$ cost;

(iii) further propose **B2P-U**, which is implemented via sampled incomplete U-statistics, with a reservoir-based stabilization to further stabilize moment estimation while remaining $O(B)$ cost.

Theoretically, we quantify the $O(B^{-1})$ bias of naive mini-batch PCC/CCC gradients, establish a tracking theorem for B2P-EMA, and provide an idealized control-variate analysis for B2P-U. We empirically show on synthetic and real-world benchmarks that B2P-Corr stabilizes training and improves correlation with negligible overhead.

2 Related Work

Correlation Objectives in Deep Learning Correlation-based losses are common when evaluation emphasizes agreement in shape or calibration. In affective computing and emotion recognition, CCC is a standard metric and is often optimized directly, typically improving over MSE but also requiring MSE warmup or careful schedules for stability [2, 13, 27]. In multi-view representation learning, deep CCA similarly depends on reliable covariance estimates and often requires large batches or additional regularization [1, 10]. Since large batches can be impractical and may hurt generalization [7, 8], stable small-batch correlation optimization remains an important practical problem.

Optimization of Non-decomposable Objectives Most work on non-decomposable stochastic optimization focuses on classification metrics such as F-measure, AUC, and AUPRC. Beyond primal–dual and saddle-point methods for confusion-matrix ratios [19, 14, 5, 6, 26], recent work studies compositional/coupled formulations and other gradient-based surrogates, including SCGD for compositions of expectations [24], finite-sum coupled compositional optimization for AP and related objectives [23], AUPRC optimization via dependent compositional objectives [15], and pAUC maximization via DRO-based methods [28]. Incomplete U-statistics provide another $O(B)$ route for pairwise objectives [4, 11]. Our method is related in spirit but targets PCC/CCC regression, using global moment sketches and stop-gradient pseudo-losses to remain within a standard SGD pipeline.

3 Correlation Objectives and B2P-Corr Decomposition

3.1 Problem Setup and Global Moments

We consider supervised regression with data $(x, y) \sim \mathcal{D}$, where $x \in \mathcal{X}$ and $y \in \mathbb{R}$ (straightforward to extend to vector outputs). A model $f_\theta : \mathcal{X} \to \mathbb{R}$ produces predictions $f_\theta(x)$ with parameters $\theta \in \mathbb{R}^d$.

We denote batch size with B and the population moments of targets and predictions by

$$\mu_y = \mathbb{E}[y], \quad \mu_f = \mathbb{E}[f_\theta(x)], \quad \sigma_y^2 = \mathrm{Var}(y), \quad \sigma_f^2 = \mathrm{Var}(f_\theta(x)), \tag{1}$$

$$C = \mathrm{Cov}(y, f_\theta(x)) = \mathbb{E}[(y - \mu_y)(f_\theta(x) - \mu_f)]\,. \tag{2}$$

Throughout, σ_y^2 and μ_y are independent of θ, while (μ_f, σ_f^2, C) depend on θ through f_θ.

The population PCC and CCC between $f_\theta(x)$ and y can be expressed as

$$\rho(\theta) = \frac{\mathrm{Cov}(y, f_\theta(x))}{\sigma_y \sigma_f} = \frac{C}{\sigma_y \sigma_f}, \;\; \mathrm{CCC}(\theta) = \frac{2C}{\sigma_y^2 + \sigma_f^2 + (\mu_y - \mu_f)^2} = \frac{2C}{D}, \tag{3}$$

where $D := \sigma_y^2 + \sigma_f^2 + (\mu_y - \mu_f)^2$. Both objectives are *non-decomposable*: they depend on global moments (μ_f, σ_f^2, C), which are unknown and must be estimated from data.

Stop-gradient Operator. We use $\mathrm{sg}(\cdot)$ to denote a stop-gradient operator: the value of $\mathrm{sg}(u)$ is equal to u in the forward pass, but $\partial\, \mathrm{sg}(u)/\partial u = 0$ in backprop. In autodiff frameworks (e.g., PyTorch [12]), this corresponds to `detach` or `stop_gradient`. This will let us treat moment sketches as constants when differentiating pseudo-losses.

3.2 Population Gradient Decompositions

Although PCC/CCC are non-decomposable as scalar objectives, we show their *population gradients* admit a clean per-example decomposition where only a few global moments appear as coefficients. These identities are the basis of the B2P-Corr framework.

Lemma 3.1 (Population gradient of PCC)**.** *Let $\rho(\theta) = C/(\sigma_y \sigma_f)$ with moments defined above. Then*

$$\nabla_\theta \rho = \mathbb{E}\left[\left(\frac{y - \mu_y}{\sigma_y \sigma_f} - \rho\, \frac{f_\theta(x) - \mu_f}{\sigma_f^2}\right) \nabla_\theta f_\theta(x)\right].$$

Lemma 3.2 (Population gradient of CCC)**.** *Recall $\mathrm{CCC}(\theta) = 2C/D$ with C and D defined above, then*

$$\nabla_\theta \mathrm{CCC} = \mathbb{E}\left[\frac{2}{D^2}\Big(D(y - \mu_y) - 2C(f_\theta(x) - \mu_y)\Big) \nabla_\theta f_\theta(x)\right].$$

Proofs are shown in Sections B.2 and B.3. Lemma 3.1 and 3.2 show that, if we can provide accurate estimates of the global moments, then correlation learning reduces to optimizing an expectation of *per-example* weighted gradients.

3.3 B2P Decomposition via Stop-gradient Pseudo-losses

For scalar predictions $f = f_\theta(x)$ and target y, consider the pseudo-losses

$$\mathcal{L}_\rho(f; y) = -\frac{y - \mathrm{sg}(\mu_y)}{\mathrm{sg}(\sigma_y)\, \mathrm{sg}(\sigma_f)} f + \frac{\mathrm{sg}(\rho)}{2\, \mathrm{sg}(\sigma_f)^2} \big(f - \mathrm{sg}(\mu_f)\big)^2, \tag{4}$$

$$\mathcal{L}_{\mathrm{CCC}}(f; y) = -\frac{2}{\mathrm{sg}(D)^2}\Big(\mathrm{sg}(D)(y - \mathrm{sg}(\mu_y)) f - \mathrm{sg}(C)\big(f - \mathrm{sg}(\mu_y)\big)^2\Big). \tag{5}$$

Note that these pseudo-losses need not equal $-\rho$ or $-\mathrm{CCC}$, we only require its gradient to match the target population gradient when the moments are correct. Treating the moments inside $\mathrm{sg}(\cdot)$ as constants, the derivative w.r.t. f gives

$$\frac{\partial \mathcal{L}_\rho}{\partial f} = -\frac{y - \mu_y}{\sigma_y \sigma_f} + \rho\, \frac{f - \mu_f}{\sigma_f^2}, \quad \frac{\partial \mathcal{L}_{\mathrm{CCC}}}{\partial f} = -\frac{2}{D^2}\Big(D(y - \mu_y) - 2C(f - \mu_y)\Big). \tag{6}$$

Therefore backpropagation yields $\mathbb{E}[\nabla_\theta \mathcal{L}_\rho] = -\nabla_\theta \rho$ (negative because we maximize PCC/CCC) and $\mathbb{E}[\nabla_\theta \mathcal{L}_{\mathrm{CCC}}] = -\nabla_\theta \mathrm{CCC}$ provided we plug in the true moments. In standard offline setting, (μ_y, σ_y^2) can be precomputed from training labels. In online settings, we simply maintain a EMA sketch as $(\hat{\mu}_y, \widehat{\sigma_y^2})$ since these moments are θ-independent and will not alter the gradient decompositions.

Now, pseudo-losses in Equations (4) and (5) are decomposable given the moments, but in practice we do not know the true moments. In next section, we introduce our B2P-Corr which maintains lightweight "moment sketches" with two $O(B)$ sketching strategies, B2P-EMA and B2P-U.

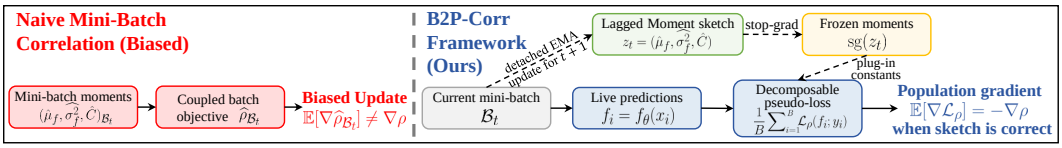


Figure 1: Batch-to-population mismatch in naive mini-batch and our proposed B2P-Corr fix.

4 Method: B2P-Corr

We propose B2P-Corr, a simple drop-in framework for mini-batch optimization of PCC/CCC which (i) use the decomposable pseudo-losses in Equations (4) and (5), and (ii) maintain a small "moment sketch" of the required global moments, treated as constants using $\mathrm{sg}(\cdot)$. The remaining design choice is how to update the sketch from mini-batches. We provide two practical instantiations: **B2P-EMA** a running-moment sketch and **B2P-U** a variance-reduced replacement for the covariance/variance estimates used inside B2P-EMA. Both share the same pseudo-loss.

4.1 B2P-EMA: Exponential-moving Moment Sketch

B2P-EMA maintains exponentially weighted average (EMA) estimates of the prediction moments $(\hat{\mu}_f, \widehat{\sigma_f^2}, \hat{C})$. On each batch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{B}$, compute predictions $f_i = f_\theta(x_i)$ and define $\tilde{f}_i = \mathrm{sg}(f_i)$ for moment updates. At iteration t, B2P-EMA computes

$$\overline{f}_{\mathcal{B}} = \frac{1}{B}\sum_i \tilde{f}_i, \quad \mathrm{var}_{\mathcal{B}}(f) = \frac{1}{B-1}\sum_i (\tilde{f}_i - \overline{f}_{\mathcal{B}})^2, \quad \mathrm{cov}_{\mathcal{B}}(y, f) = \frac{1}{B-1}\sum_i (y_i - \mu_y)(\tilde{f}_i - \overline{f}_{\mathcal{B}}), \tag{7}$$

and updates

$$\hat{\mu}_f \leftarrow (1-\alpha_t)\hat{\mu}_f + \alpha_t \overline{f}_{\mathcal{B}}, \quad \widehat{\sigma_f^2} \leftarrow (1-\alpha_t)\widehat{\sigma_f^2} + \alpha_t \mathrm{var}_{\mathcal{B}}(f), \quad \hat{C} \leftarrow (1-\alpha_t)\hat{C} + \alpha_t \,\mathrm{cov}_{\mathcal{B}}(y, f), \tag{8}$$

then form derived quantities (with $\varepsilon, \delta > 0$ for stability)

$$\hat{\sigma}_f = \sqrt{\widehat{\sigma_f^2} + \varepsilon}, \quad \hat{\rho} = \mathrm{clip}\left(\frac{\hat{C}}{\sigma_y \hat{\sigma}_f}, -1+\delta, 1-\delta\right), \quad \hat{D} = \sigma_y^2 + \widehat{\sigma_f^2} + (\mu_y - \hat{\mu}_f)^2. \tag{9}$$

Finally, optimize the decomposable batch objective $\frac{1}{B}\sum_i \mathcal{L}_\rho(f_i; y_i)$ or $\frac{1}{B}\sum_i \mathcal{L}_{\mathrm{CCC}}(f_i; y_i)$ where all estimated moments appear under $\mathrm{sg}(\cdot)$. Algorithm 1 summarizes one step and the extra cost over naive mini-batch training is $O(B)$ update with negligible overhead.

Remark. Theorem 5.3 uses a diminishing α_t for a clean asymptotic tracking statement. In experiments we use a constant α, which corresponds to keeping a small steady-state tracking error when the target moments drift slowly.

4.2 B2P-U: Stabilized Plug-in Sketch

While B2P-EMA greatly stabilizes correlation optimization, $\mathrm{var}_{\mathcal{B}}(f)$ and $\mathrm{cov}_{\mathcal{B}}(y, f)$ still fluctuate under small B. B2P-U further stabilizes these moment estimates using cross-batch U-statistic identities together with a small reservoir-based correction.

Cross-batch U-statistic identities. For i.i.d. pairs $((x_1, y_1), (x_2, y_2))$ and predictions $f_i = f_\theta(x_i)$, the moments satisfy

$$\mathrm{Cov}(y, f) = \tfrac{1}{2}\,\mathbb{E}\big[(y_1 - y_2)(f_1 - f_2)\big], \quad \mathrm{Var}(f) = \tfrac{1}{2}\,\mathbb{E}\big[(f_1 - f_2)^2\big]. \tag{10}$$

with proof in Section B.1. For the second-order moments considered here, the complete cross-batch estimator over all B^2 pairs also admits an exact algebraically collapsed $O(B)$ form. In our implementation we instead use a sampled incomplete U-statistic with $K = O(B)$ cross-batch pairs, which preserves the same pairwise identity while fitting naturally with our cross-batch sketch update. To reduce coupling between the sketch (in denominators) and the current per-example gradients, we form pairs across consecutive batches: at step t, sample cross-batch pairs $(y_i, \tilde{f}_i)$ and $(y'_j, \tilde{f}'_j)$ with $i \in \mathcal{B}_t$, $j \in \mathcal{B}_{t-1}$ and compute

$$\hat{C}_{\mathrm{U}} = \frac{1}{2K}\sum_{(i,j)} (y_i - y'_j)(\tilde{f}_i - \tilde{f}'_j), \quad \widehat{\sigma^2_{f,\mathrm{U}}} = \frac{1}{2K}\sum_{(i,j)} (\tilde{f}_i - \tilde{f}'_j)^2. \tag{11}$$

Algorithm 1 B2P-EMA training step for PCC/CCC

Input: mini-batch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{B\geq 2}$; moment sketch $(\hat{\mu}_f, \widehat{\sigma_f^2}, \hat{C})$; label moments (μ_y, σ_y); stability $\varepsilon, \delta > 0$; EMA rate $\alpha_t \in (0, 1]$
$f_i \leftarrow f_\theta(x_i)$ for all $i \in \mathcal{B}$; $\tilde{f}_i \leftarrow \mathrm{sg}(f_i)$
$\overline{f}_{\mathcal{B}} \leftarrow \frac{1}{B}\sum_{i=1}^{B} \tilde{f}_i$; $\mathrm{var}_{\mathcal{B}}(f) \leftarrow \frac{1}{B-1}\sum_{i=1}^{B}(\tilde{f}_i - \overline{f}_{\mathcal{B}})^2$; $\mathrm{cov}_{\mathcal{B}}(y, f) \leftarrow \frac{1}{B-1}\sum_{i=1}^{B}(y_i - \mu_y)(\tilde{f}_i - \overline{f}_{\mathcal{B}})$
$\hat{\mu}_f \leftarrow (1-\alpha_t)\hat{\mu}_f + \alpha_t \overline{f}_{\mathcal{B}}$; $\widehat{\sigma_f^2} \leftarrow (1-\alpha_t)\widehat{\sigma_f^2} + \alpha_t \mathrm{var}_{\mathcal{B}}(f)$; $\hat{C} \leftarrow (1-\alpha_t)\hat{C} + \alpha_t\, \mathrm{cov}_{\mathcal{B}}(y, f)$
$\hat{\sigma}_f \leftarrow \sqrt{\widehat{\sigma_f^2} + \varepsilon}$; $\hat{\rho} \leftarrow \mathrm{clip}\big(\hat{C}/(\sigma_y \hat{\sigma}_f), -1+\delta, 1-\delta\big)$; $\hat{D} \leftarrow \sigma_y^2 + \widehat{\sigma_f^2} + (\mu_y - \hat{\mu}_f)^2$ ▷ CCC
$\mathcal{L} \leftarrow \frac{1}{B}\sum_{i=1}^{B} \mathcal{L}(f_i; y_i)$ using Equation (4) (use Equation (5) if loss = CCC)
Compute $\nabla_\theta \mathcal{L}$ and update θ with an optimizer step

Reservoir-based stabilization. We maintain a small reservoir $\mathcal{R}$ (size $m \ll |\mathcal{D}|$) of cached $\tilde{f}$ and labels y (refreshed periodically). For any moment $M \in \{C, \sigma_f^2\}$, we combine the sampled U-statistic estimate $\hat{M}_{\mathrm{U}}$ with a reservoir estimate $\hat{M}_{\mathcal{R}}$ via

$$\hat{M}^\star = \hat{M}_{\mathrm{U}} + \lambda\big(\hat{M}_{\mathcal{R}} - \widehat{\mathbb{E}}[\hat{M}_{\mathcal{R}}]\big),$$

where $\widehat{\mathbb{E}}[\hat{M}_{\mathcal{R}}]$ is a running mean of $\hat{M}_{\mathcal{R}}$ tracked by an EMA. In the classical idealized control-variate setting, the optimal choice is

$$\lambda^\star = -\frac{\mathrm{Cov}(\hat{M}_{\mathrm{U}}, \hat{M}_{\mathcal{R}})}{\mathrm{Var}(\hat{M}_{\mathcal{R}})}. \tag{12}$$

In our implementation, however, the reservoir is disjoint and periodically refreshed, so the EMA-estimated coefficient should be viewed as a practical stabilization heuristic rather than a guarantee of stepwise conditional variance reduction at fixed θ_t; it can also reflect temporal correlation induced by slow parameter drift. We then update $(\hat{C}, \widehat{\sigma_f^2})$ in Equation (8) using $\hat{C}^\star$ and $\widehat{\sigma_f^2}^\star$ in place of the raw mini-batch estimators. Algorithm 2 details the B2P-U subroutine, which preserves $O(B)$ complexity in our implementation.

5 Theoretical Study

We analyze three aspects of correlation optimization: (i) the bias of naive mini-batch PCC/CCC gradients; (ii) moving-target tracking of the B2P-EMA sketch and the induced plug-in gradient; and (iii) an idealized control-variate analysis that motivates the reservoir stabilization used in B2P-U.

Assumption 5.1 (Differentiability, Integrability, and Non-degeneracy)**.** Fix θ and write $f := f_\theta(x)$. Define the raw-moment vector $s := (f, f^2, yf) \in \mathbb{R}^3$, and its Jacobian $\nabla_\theta s \in \mathbb{R}^{3\times d}$. Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^3$ and $\|\cdot\|_F$ the Frobenius norm on $\mathbb{R}^{3\times d}$. Assume:

(1) **Interchanging ∇_θ and $\mathbb{E}$.** $f_\theta(x)$ is a.s. differentiable in θ and for each coordinate of s we may interchange differentiation and expectation, e.g., $\nabla_\theta \mathbb{E}[f] = \mathbb{E}[\nabla_\theta f]$, $\nabla_\theta \mathbb{E}[f^2] = \mathbb{E}[\nabla_\theta (f^2)]$, and $\nabla_\theta \mathbb{E}[yf] = \mathbb{E}[y \nabla_\theta f]$.

(2) **Finite moments.** $\mathbb{E}[y^2] < \infty$, $\mathbb{E}[\|s\|^2] < \infty$, $\mathbb{E}[\|\nabla_\theta s\|_F^2] < \infty$, and $\mathbb{E}[\|s\|^2 \, \|\nabla_\theta s\|_F] < \infty$.

(3) **Non-degeneracy.** $\sigma_y > 0$ and either (i) $\sigma_f(\theta) \geq \sigma_{\min} > 0$, or (ii) we analyze ε-stabilized versions with $\sigma_f^2 \leftarrow \sigma_f^2 + \varepsilon$ and $D \leftarrow D + \varepsilon$.

5.1 Bias of Naive Mini-batch PCC/CCC Gradients

Let $\widehat{\rho}_{\mathcal{B}}$ and $\widehat{\mathrm{CCC}}_{\mathcal{B}}$ denote PCC and CCC computed from a random i.i.d. mini-batch $\mathcal{B}$ of size $B \geq 2$, where the moments inside the ratio are computed from the same batch and we differentiate through those batch moments. This is the standard "naive" mini-batch training strategy.

Proposition 5.2 (Naive Mini-batch Gradients Have $O(1/B)$ Bias)**.** *Under Assumption 5.1, for i.i.d. mini-batches of size B we have*

$$\mathbb{E}_{\mathcal{B}}\big[\nabla_\theta \widehat{\rho}_{\mathcal{B}}\big] = \nabla_\theta \rho + O(B^{-1}), \quad \mathbb{E}_{\mathcal{B}}\big[\nabla_\theta \widehat{\mathrm{CCC}}_{\mathcal{B}}\big] = \nabla_\theta \mathrm{CCC} + O(B^{-1}), \tag{13}$$

Algorithm 2 B2P-U sketch update plug-in for Algorithm 1

Input: current batch $(y_i, \tilde{f}_i)_{i=1}^B$ and previous batch $(y'_j, \tilde{f}'_j)_{j=1}^B$; reservoir $\mathcal{R}$ (size m) storing cached pairs $(y, \tilde{f})$; current sketch $(\hat{C}, \widehat{\sigma_f^2})$; EMA rate α_t; number of pairs $K = O(B)$; $\tau > 0$

Sample K cross-batch pairs (i, j) with $i, j \in \{1, \dots, B\}$

$\hat{C}_{\mathrm{U}} \leftarrow \frac{1}{2K} \sum_{(i,j)} (y_i - y'_j)\,(\tilde{f}_i - \tilde{f}'_j)$; $\widehat{\sigma^2_{f,\mathrm{U}}} \leftarrow \frac{1}{2K} \sum_{(i,j)} (\tilde{f}_i - \tilde{f}'_j)^2$

Update running means $\widehat{\mathbb{E}}[\hat{C}_{\mathcal{R}}]$, $\widehat{\mathbb{E}}[\widehat{\sigma^2_{f,\mathcal{R}}}]$ (e.g., via EMA)

Update running estimates of $\mathrm{Var}(\hat{C}_{\mathcal{R}})$, $\mathrm{Cov}(\hat{C}_{\mathrm{U}}, \hat{C}_{\mathcal{R}})$ and $\mathrm{Var}(\widehat{\sigma^2_{f,\mathcal{R}}})$, $\mathrm{Cov}(\widehat{\sigma^2_{f,\mathrm{U}}}, \widehat{\sigma^2_{f,\mathcal{R}}})$

$\lambda_C \leftarrow -\mathrm{Cov}(\hat{C}_{\mathrm{U}}, \hat{C}_{\mathcal{R}}) / \big(\mathrm{Var}(\hat{C}_{\mathcal{R}}) + \tau\big)$; $\lambda_V \leftarrow -\mathrm{Cov}(\widehat{\sigma^2_{f,\mathrm{U}}}, \widehat{\sigma^2_{f,\mathcal{R}}}) / \big(\mathrm{Var}(\widehat{\sigma^2_{f,\mathcal{R}}}) + \tau\big)$

$\hat{C}^\star \leftarrow \hat{C}_{\mathrm{U}} + \lambda_C \big(\hat{C}_{\mathcal{R}} - \widehat{\mathbb{E}}[\hat{C}_{\mathcal{R}}]\big)$; $\widehat{\sigma_f^{2\star}} \leftarrow \widehat{\sigma^2_{f,\mathrm{U}}} + \lambda_V \big(\widehat{\sigma^2_{f,\mathcal{R}}} - \widehat{\mathbb{E}}[\widehat{\sigma^2_{f,\mathcal{R}}}]\big)$

$\hat{C} \leftarrow (1 - \alpha_t)\hat{C} + \alpha_t \hat{C}^\star$; $\widehat{\sigma_f^2} \leftarrow (1 - \alpha_t)\widehat{\sigma_f^2} + \alpha_t \widehat{\sigma_f^{2\star}}$ ▷ Reduced-variance sketch

Proceed with lines 5–7 in Algorithm 1

where the $O(B^{-1})$ term is generally nonzero due to coupling between random batch moments (appearing in denominators) and per-example gradient terms. A full proof is given in Appendix C.2. An empirical verification with synthetic dataset is given in Section 6.1.2.

This bias is the analytic expression of the B2P gap: naive mini-batch optimization directly follows the gradient of a *random* local objective whose expectation does not equal the desired population objective, and the discrepancy decreases only slowly as B increases.

5.2 Tracking of B2P-EMA

In practice, the parameters and the sketch are updated jointly, so the key question is whether the EMA sketch tracks the *current* population moments along training. Let $m(\theta) := (\mu_f(\theta), \sigma_f^2(\theta), C(\theta))$, $z_t := (\hat{\mu}_f^{(t)}, \widehat{\sigma_f^{2(t)}}, \hat{C}^{(t)})$, and let $\mathcal{F}_t$ denote all information available before drawing the batch at step t. This is the standard fast-sketch / slow-parameter regime in stochastic compositional optimization [24, 23]. We work under a standard tracking regime: the detached batch-moment estimator is conditionally unbiased with bounded second moment, the EMA steps satisfy standard diminishing-step conditions, the target moments drift slowly relative to α_t, and the mean plug-in gradient is Lipschitz in the sketch. The formal assumptions are deferred to Appendix C.3.

Theorem 5.3 (B2P-EMA tracks moving population moments)**.** *Let $J(\theta)$ denote the negative PCC objective $-\rho(\theta)$ or the negative CCC objective $-\mathrm{CCC}(\theta)$, and let g_t be the mini-batch gradient produced by the B2P-EMA pseudo-loss at step t. Under Assumption 5.1 and C.2,*

$$\mathbb{E}\|z_t - m(\theta_t)\|^2 \to 0, \quad \mathbb{E}\big\|\mathbb{E}[g_t \mid \mathcal{F}_t] - \nabla J(\theta_t)\big\| \to 0.$$

A sufficient condition for slow drift is that $m(\theta)$ is Lipschitz and $\|\theta_{t+1} - \theta_t\| = o(\alpha_t)$. Proof is given in Appendix C.4. This is a tracking result for the sketch and induced plug-in gradient, not a full convergence theorem for the coupled nonconvex training dynamics.

5.3 Idealized Control-Variate Analysis for B2P-U

Proposition 5.4 (Idealized control-variate analysis)**.** *Let $\hat{M}_U$ be an (incomplete) U-statistic estimate of a scalar moment M, and let $\hat{M}_{\mathcal{R}}$ be another estimator with $\mathbb{E}[\hat{M}_{\mathcal{R}}] = M$ and $\mathrm{Var}(\hat{M}_{\mathcal{R}}) > 0$. For any scalar λ, define $\hat{M}^\star = \hat{M}_U + \lambda\big(\hat{M}_{\mathcal{R}} - \mathbb{E}[\hat{M}_{\mathcal{R}}]\big)$. Then $\hat{M}^\star$ is unbiased for M, and choosing $\lambda^\star = -\mathrm{Cov}(\hat{M}_U, \hat{M}_{\mathcal{R}})/\mathrm{Var}(\hat{M}_{\mathcal{R}})$ yields*

$$\mathrm{Var}(\hat{M}^\star) = \mathrm{Var}(\hat{M}_U) - \frac{\mathrm{Cov}(\hat{M}_U, \hat{M}_{\mathcal{R}})^2}{\mathrm{Var}(\hat{M}_{\mathcal{R}})} \le \mathrm{Var}(\hat{M}_U),$$

with strict inequality whenever $\mathrm{Cov}(\hat{M}_U, \hat{M}_{\mathcal{R}}) \neq 0$ (proof is provided in Section C.5).

Remark 5.5 (Implementation). Proposition 5.4 is an idealized statement: it applies when the auxiliary estimator is correlated with $\hat{M}_{\mathrm{U}}$ at the target parameter. In Algorithm 2, the reservoir is disjoint and

periodically refreshed, so the classical conditional sampling covariance at fixed θ_t can be small and, under independent-sample idealizations, may vanish. Accordingly, the EMA-estimated coefficient in B2P-U should be interpreted as a practical stabilization mechanism—potentially reflecting temporal correlation induced by slowly drifting parameters—rather than a guarantee of stepwise conditional variance reduction. Empirically, B2P-U nonetheless exhibits lower gradient variability in Figure 2.

6 Experiments

We first use two synthetic experiments to verify the $O(B^{-1})$ bias of naive mini-batch PCC/CCC gradients and to compare gradient alignment/stability. We then evaluate B2P-EMA/B2P-U as drop-in replacements for naive mini-batch PCC/CCC optimization on three real-world benchmarks, followed by overhead and sensitivity studies.

6.1 Synthetic Dataset for Theory Verification

We use two controlled regression problems with a linear predictor $f_\theta(x) = \theta^\top x$ to demonstrate the "batch-to-population" (B2P) gap for correlation objectives. Unless stated otherwise, we maximize PCC $\rho(\theta)$ defined in Section 3. CCC diagnostics are reported in Figures 2c and 2d (CCC results for Exp.2 are provided in Section E.2). Experiment details and metrics are deferred to Section D.

6.1.1 Experiment Settings

Exp. 1: Imbalanced Mixture Regression. We construct an imbalanced two-component Gaussian mixture to make mini-batch moment estimates highly unstable under small batch-size B. We sample a component indicator $s \in \{+1, -1\}$ with $\mathbb{P}(s = +1) \gg \mathbb{P}(s = -1)$, draw $x \sim \mathcal{N}(s\mu, I)$, and set $y = {w^\star}^\top x + \Delta\, s + \varepsilon_s$, where Δ is a mixture-component label shift and ε_s is zero-mean heteroscedastic noise. This setting is challenging as the population moments are well-defined, but within-batch estimates of $\mathrm{Var}(f_\theta)$ and $\mathrm{Cov}(y, f_\theta)$ vary widely across batches.

Metric. At each training checkpoint with θ, we compute an oracle population gradient $g_{\mathrm{pop}}(\theta) = \nabla_\theta \rho(\theta)$ using full batch, and compare it with a stochastic gradient estimate $g(\theta)$ from a mini-batch of size B. We report (i) gradient cosine *alignment* $\cos(g(\theta), g_{\mathrm{pop}}(\theta))$, and (ii) ***RelBias***, the relative ℓ_2 error $\|\mathbb{E}[g(\theta)] - g_{\mathrm{pop}}(\theta)\|_2 / \|g_{\mathrm{pop}}(\theta)\|_2$.

Exp. 2: Isotropic Linear Regression. We sample $x \sim \mathcal{N}(0, I)$ and let $y = {w^\star}^\top x + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and $\|w^\star\|_2 = 1$. Since PCC is scale-invariant in f_θ, the optimum is characterized by directional alignment between θ and $w^\star$ rather than the norm of θ.

Metric. We train with varying batch sizes and evaluate (i) ***DirectionRecovery*** $= \cos(\theta, w^\star)$ and (ii) ***TestPCC*** = $\rho(\theta)$ on test set.

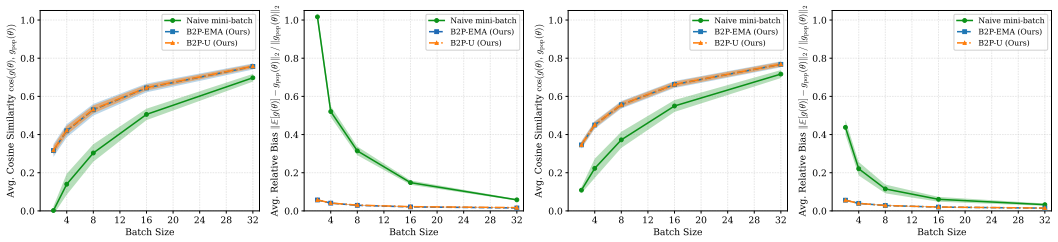
6.1.2 Result Analysis

Figures 2b and 2d empirically verifies Proposition 5.2: the average relative bias of naive PCC/CCC gradients decreases roughly linearly with $1/B$. Beyond this bias law, Figure 2 shows that B2P-Corr yields much better-aligned stochastic gradients, especially at very small batch sizes. For B=2, naive PCC gradients are often NaN or nearly random, whereas B2P remains stable; for B=4, alignment improves from 0.16 to 0.43 and RelBias drops from 0.98 to 0.08. B2P-U matches B2P-EMA in mean alignment while reduces gradient estimation variance (different from variance in Table 1). Empirically, B2P-U in Figure 2a exhibits a narrower variance shade, with average standard deviations across batch sizes dropping from 0.0346 (naive) and 0.0275 (B2P-EMA) to 0.0183 (B2P-U).

Figure 3 shows the same optimization effect in isotropic regression. With B=2, naive mini-batch PCC training fails, whereas both B2P variants recover the correct direction and substantially higher test PCC; the gap narrows as B increases, consistent with improved concentration of batch moments.

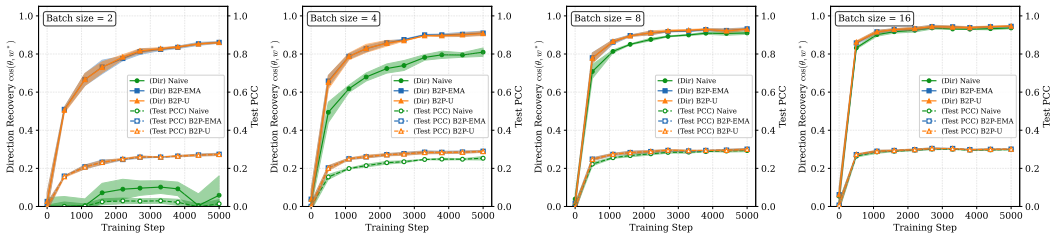
6.2 Real-world Regression Benchmarks

We evaluate B2P-Corr on three real-world tasks where the state-of-the-art baselines *directly optimize* PCC/CCC with mini-batches. We adopt a strict drop-in protocol: for each baseline, we keep the



(a) PCC: gradient align (b) PCC: relative bias (c) CCC: gradient align (d) CCC: relative bias

Figure 2: For Exp. 1, we compare stochastic gradients from naive mini-batch optimization and B2P-Corr against an oracle population gradient from full-batch training. (a–d) Average gradient cosine alignment and relative bias at different batch size B for PCC and CCC.



(a) PCC: training ($B = 2$) (b) PCC: training ($B = 4$) (c) PCC: training ($B = 8$) (d) PCC: training ($B = 16$)

Figure 3: For Exp. 2, we show DirectionRecovery and TestPCC vs. training steps for different batch sizes. With B=2, naive mini-batch fails to converge while B2P-EMA/B2P-U continue to improve. The gap narrows as B increases. CCC training curves are shown in Section E.2.

architecture, data splits, and auxiliary losses unchanged, and only replace the naive mini-batch correlation objective with our decomposable B2P pseudo-loss (B2P-EMA or B2P-U) and a fine-grained PCC loss alternative PLC [3]. Dataset details and training protocols are provided in Section D.

6.2.1 Datasets

Video Quality Assessment (PCC). KoNViD-1k comprises 1,200 in-the-wild videos (960×540) sampled from YFCC100M [22]. Following KVQ [16], we use Mean Opinion Scores (MOS) as targets and split 960/120/120 videos for train/val/test. KVQ directly optimizes PCC and reports PCC, SROCC, and RMSE.

Spatial Transcriptomic (PCC). The 10xProteomic dataset used by EGN [25] contains 32,032 pathology image patches paired with gene-expression vectors $\mathbf{y}_i \in \mathbb{R}_+^{|G|}$ for breast cancer. We follow its data preprocessing and the 3-fold evaluation. EGN jointly optimizes MSE+PCC during training and reports PCC, MAE, and MSE.

Video-based Sentiment Analysis (CCC). VEATIC contains 124 videos (257,601 frames) with valence $y_i^{(v)} \in \mathbb{R}$ and arousal $y_i^{(a)} \in \mathbb{R}$ annotation. The baseline optimizes joint learning objective CCC+MSE and we report CCC, PCC, and RMSE.

6.2.2 Result Analysis

Table 1 shows that replacing the naive mini-batch PCC/CCC term with B2P-EMA/B2P-U consistently improves correlation across all three tasks and batch sizes, while keeping architectures, optimizers, preprocessing, and auxiliary losses unchanged.

Video Quality Assessment (KoNViD-1k). In the unstable small-batch regime (B=2), B2P-EMA yields substantial gains, boosting PCC from 0.865 to 0.895 and reducing RMSE by 6.4% (0.303 → 0.283). This significantly outperforms the fine-grain correlation loss PLC (0.871), confirming that the primary bottleneck is the bias in local moment estimation rather than model capacity. As batch size increases, B2P-Corr variants maintain superiority, with B2P-U achieving the best overall performance at $B = 8$ (PCC 0.906, SROCC 0.893, RMSE 0.264). This aligns with our analysis that naive mini-batch correlation gradients couple per-example terms with denominators (Proposition 5.2), and our B2P-U effectiveness reduces variance during moment estimation.

Table 1: Comparison across three datasets. We replace naive mini-batch PCC/CCC objectives in each baseline with our B2P-Corr and it consistently improves performance. "Default" denotes the original batch size used in each baseline. The VEATIC arousal prediction result are shown in Section E.1.

Dataset	Batch-size	2 (Default)			4			8		
	Metrics	PCC ↑	SROCC ↑	RMSE ↓	PCC ↑	SROCC ↑	RMSE ↓	PCC ↑	SROCC ↑	RMSE ↓
KoNViD-1k (PCC)	KVQ [16]	$0.865_{\pm 0.028}$	$0.861_{\pm 0.023}$	$0.303_{\pm 0.018}$	$0.878_{\pm 0.029}$	$0.863_{\pm 0.032}$	$0.299_{\pm 0.017}$	$0.881_{\pm 0.033}$	$0.875_{\pm 0.026}$	$0.290_{\pm 0.015}$
	+PLC [3]	$0.871_{\pm 0.033}$	$0.864_{\pm 0.027}$	$0.301_{\pm 0.016}$	$0.881_{\pm 0.031}$	$0.853_{\pm 0.021}$	$0.289_{\pm 0.010}$	$0.887_{\pm 0.028}$	$0.871_{\pm 0.035}$	$0.293_{\pm 0.015}$
	+B2P-EMA	$\underline{0.895}_{\pm 0.024}$	$\mathbf{0.880}_{\pm 0.030}$	$\mathbf{0.283}_{\pm 0.011}$	$\underline{0.893}_{\pm 0.023}$	$\mathbf{0.884}_{\pm 0.019}$	$\underline{0.279}_{\pm 0.017}$	$\underline{0.895}_{\pm 0.026}$	$\underline{0.881}_{\pm 0.022}$	$\underline{0.280}_{\pm 0.010}$
	+B2P-U	$\underline{0.887}_{\pm 0.024}$	$\underline{0.867}_{\pm 0.018}$	$\underline{0.289}_{\pm 0.013}$	$\mathbf{0.896}_{\pm 0.028}$	$\underline{0.881}_{\pm 0.021}$	$\mathbf{0.278}_{\pm 0.013}$	$\mathbf{0.906}_{\pm 0.025}$	$\mathbf{0.893}_{\pm 0.028}$	$\mathbf{0.264}_{\pm 0.010}$

Dataset	Batch-size	8			16			32 (Default)		
	Metrics	PCC ↑	MAE ↓	MSE ↓	PCC ↑	MAE ↓	MSE ↓	PCC ↑	MAE ↓	MSE ↓
10xProteomic (PCC)	EGN [25]	$0.615_{\pm 0.094}$	$0.166_{\pm 0.085}$	$0.056_{\pm 0.052}$	$0.637_{\pm 0.150}$	$\underline{0.161}_{\pm 0.091}$	$0.057_{\pm 0.059}$	$0.629_{\pm 0.150}$	$0.165_{\pm 0.098}$	$0.061_{\pm 0.066}$
	+PLC [3]	$0.623_{\pm 0.097}$	$0.165_{\pm 0.087}$	$0.055_{\pm 0.049}$	$0.645_{\pm 0.139}$	$0.162_{\pm 0.101}$	$0.059_{\pm 0.064}$	$0.631_{\pm 0.149}$	$0.169_{\pm 0.110}$	$0.061_{\pm 0.063}$
	+B2P-EMA	$\underline{0.644}_{\pm 0.089}$	$\underline{0.159}_{\pm 0.083}$	$\underline{0.052}_{\pm 0.049}$	$\mathbf{0.675}_{\pm 0.125}$	$0.162_{\pm 0.103}$	$\underline{0.056}_{\pm 0.061}$	$\underline{0.665}_{\pm 0.144}$	$0.159_{\pm 0.101}$	$0.055_{\pm 0.059}$
	+B2P-U	$\mathbf{0.660}_{\pm 0.091}$	$\mathbf{0.155}_{\pm 0.080}$	$\mathbf{0.051}_{\pm 0.048}$	$\underline{0.667}_{\pm 0.134}$	$\mathbf{0.158}_{\pm 0.093}$	$\mathbf{0.053}_{\pm 0.057}$	$\mathbf{0.673}_{\pm 0.161}$	$\mathbf{0.155}_{\pm 0.087}$	$\mathbf{0.051}_{\pm 0.073}$

Dataset	Batch-size	4			10			20 (Default)		
	Metrics	CCC ↑	PCC ↑	RMSE ↓	CCC ↑	PCC ↑	RMSE ↓	CCC ↑	PCC ↑	RMSE ↓
VEATIC Valence (CCC)	VEATIC [17]	$0.437_{\pm 0.036}$	$0.520_{\pm 0.029}$	$0.311_{\pm 0.015}$	$0.574_{\pm 0.045}$	$0.638_{\pm 0.028}$	$0.294_{\pm 0.019}$	$0.606_{\pm 0.041}$	$0.639_{\pm 0.038}$	$0.317_{\pm 0.012}$
	+B2P-EMA	$\underline{0.460}_{\pm 0.035}$	$\mathbf{0.541}_{\pm 0.037}$	$\underline{0.306}_{\pm 0.013}$	$\underline{0.585}_{\pm 0.040}$	$\underline{0.644}_{\pm 0.033}$	$\underline{0.291}_{\pm 0.015}$	$\underline{0.613}_{\pm 0.035}$	$\underline{0.660}_{\pm 0.033}$	$\mathbf{0.298}_{\pm 0.009}$
	+B2P-U	$\mathbf{0.467}_{\pm 0.025}$	$\underline{0.538}_{\pm 0.028}$	$\mathbf{0.298}_{\pm 0.010}$	$\mathbf{0.597}_{\pm 0.043}$	$\mathbf{0.655}_{\pm 0.031}$	$\mathbf{0.287}_{\pm 0.015}$	$\mathbf{0.615}_{\pm 0.026}$	$\mathbf{0.671}_{\pm 0.022}$	$\underline{0.301}_{\pm 0.008}$

Spatial Transcriptomic (10xProteomic). Table 1 reports 10xProteomic gene-expression prediction results under 3-fold evaluation. The result shows that replacing the naive PCC term in EGN's objective with B2P-Corr yields consistent improvements. In the high-variance setting (B=8), B2P-U is most effective, improving PCC ($0.615 \rightarrow 0.660$) while simultaneously reducing MAE and MSE. This indicates that stable correlation optimization enhances rather than compromises magnitude metrics. At the default B=32, B2P-U scales effectively, achieving the highest PCC (0.673) and lowest MSE. Overall, these results highlight B2P-Corr's robustness in heterogeneous, high-noise regimes where reliable global moment estimates are critical.

Video-based Sentiment Analysis (VEATIC). Table 1 also demonstrates B2P-Corr effectively improves and stabilizes CCC optimization, which is more sensitive to moment mismatch than PCC because its normalization term D couples variance and mean mismatch in addition to covariance. At B=4, B2P-U improves CCC ($0.437 \rightarrow 0.467$) and RMSE ($0.311 \rightarrow 0.298$). B2P-U remains dominant at B=10, yielding the best CCC (0.597) and lowest RMSE (0.287). Even at the default B=20, B2P-U boosts CCC from 0.606 to 0.615, verifying that decoupling the gradient calculation from noisy batch denominators prevents the estimation error propagation.

6.3 Efficiency and Sensitivity

Table 2 shows B2P-EMA adds essentially no runtime or VRAM cost, while B2P-U incurs less than 1% extra runtime from reservoir updates and cross-batch pairing with no additional VRAM. Sensitivity to α and reservoir size is reported in Sections E.3 and E.4. Both variants remain stable across practical ranges. Empirically, for B2P-EMA, larger batches tolerate larger α since batch statistics are closer to population moments. For B2P-U, a larger reservoir reduces estimation noise but increases higher runtime.

Table 2: **KoNViD-1k runtime and memory overhead.** We evaluate per-iteration runtime and VRAM usage for KVQ with different batch sizes. B2P-Corr adds negligible time and VRAM overhead, while improving test PCC relative to naive mini-batch correlation objective.

Batch-size	2 (default)		4		8	
Overhead	Sec/iter	VRAM	Sec/iter	VRAM	Sec/iter	VRAM
KVQ	3.168	25.102 GB	5.730	44.914 GB	11.061	76.216 GB
+B2P-EMA	3.169	25.102 GB	5.734	44.914 GB	11.093	76.216 GB
+B2P-U	3.177	25.102 GB	5.764	44.914 GB	11.165	76.216 GB

7 Conclusion

We introduced **B2P-Corr**, a simple framework for closing the batch-to-population gap in correlation learning with decomposable stop-gradient pseudo-losses and lightweight moment sketches. **B2P-EMA** provides asymptotic gradient tracking, while **B2P-U** adds sampled cross-batch U-statistics with reservoir-based stabilization. Across synthetic and real-world tasks, B2P-Corr reduces the batch-to-population gap, stabilizes training, and improves correlation with negligible overhead.

8 Acknowledgement

This work was partially supported by US National Science Foundation IIS-2412195, CCF-2400785, the Cancer Prevention and Research Institute of Texas (CPRIT) award (RP230363), the National Institutes of Health (NIH) R01 awards (1R01AI190103-01 and 1R01LM015497-01).

References

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

[2] Bagus Tris Atmaja and Masato Akagi. Evaluation of error-and correlation-based loss functions for multitask learning dimensional speech emotion recognition. In *Journal of Physics: Conference Series*, volume 1896, page 012004. IOP Publishing, 2021.

[3] Chaoyu Chen, Xin Yang, Ruobing Huang, Xindi Hu, Yankai Huang, Xiduo Lu, Xinrui Zhou, Mingyuan Luo, Yinyu Ye, Xue Shuang, et al. Fine-grained correlation loss for regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 663–672. Springer, 2022.

[4] Stephan Clémençon, Igor Colin, and Aurélien Bellet. Scaling-up empirical risk minimization: optimization of incomplete u-statistics. *Journal of Machine Learning Research*, 17(76):1–36, 2016.

[5] Krzysztof Dembczynski, Arkadiusz Jachnik, Wojciech Kotlowski, Willem Waegeman, and Eyke Huellermeier. Optimizing the f-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1130–1138, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `https://proceedings.mlr.press/v28/dembczynski13.html`.

[6] Wei Gao, Rong Jin, Shenghuo Zhu, and Zhi-Hua Zhou. One-pass auc optimization. In *International conference on machine learning*, pages 906–914. PMLR, 2013.

[7] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[8] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[9] Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

[10] Sehwan Moon, Jeongyoung Hwang, and Hyunju Lee. Sdgcca: supervised deep generalized canonical correlation analysis for multi-omics integration. *Journal of Computational Biology*, 29(8):892–907, 2022.

[11] Guillaume Papa, Stéphan Clémençon, and Aurélien Bellet. Sgd algorithms based on incomplete u-statistics: large-scale minimization of empirical risk. *Advances in Neural Information Processing Systems*, 28, 2015.

[12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[13] R Gnana Praveen and Jahangir Alam. Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4803–4813, 2024.

[14] Shameem Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. Optimizing f-measures by cost-sensitive classification. *Advances in neural information processing systems*, 27, 2014.

[15] Qi Qi, Youzhi Luo, Zhao Xu, Shuiwang Ji, and Tianbao Yang. Stochastic optimization of areas under precision-recall curves with provable convergence. *Advances in neural information processing systems*, 34:1752–1765, 2021.

[16] Yunpeng Qu, Kun Yuan, Qizhi Xie, Ming Sun, Chao Zhou, and Jian Wang. Kvq: Boosting video quality assessment via saliency-guided local perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2150–2160, 2025.

[17] Zhihang Ren, Jefferson Ortega, Yifan Wang, Zhimin Chen, Yunhui Guo, Stella X Yu, and David Whitney. Veatic: Video-based emotion and affect tracking in context dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4467–4477, 2024.

[18] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[19] Amartya Sanyal, Pawan Kumar, Purushottam Kar, Sanjay Chawla, and Fabrizio Sebastiani. Optimizing non-decomposable measures with deep networks. *Machine Learning*, 107(8): 1597–1620, 2018.

[20] Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768, 2018.

[21] Michael A Skinnider, Jordan W Squair, and Leonard J Foster. Evaluating measures of association for single-cell transcriptomics. *Nature methods*, 16(5):381–386, 2019.

[22] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[23] Bokun Wang and Tianbao Yang. Finite-sum coupled compositional stochastic optimization: Theory and applications. *arXiv preprint arXiv:2202.12396*, 2022.

[24] Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1):419–449, 2017.

[25] Yan Yang, Md Zakir Hossain, Eric A Stone, and Shafin Rahman. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5039–5048, 2023.

[26] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3040–3049, 2021.

[27] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan. Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2376–2381, 2022.

[28] Dixian Zhu, Gang Li, Bokun Wang, Xiaodong Wu, and Tianbao Yang. When auc meets dro: Optimizing partial auc for deep learning with non-convex convergence guarantee. In *International Conference on Machine Learning*, pages 27548–27573. PMLR, 2022.

Appendix

A Reproducibility

We provide a PyTorch [12] implementation of B2P-EMA and B2P-U in Github. B2P-EMA and B2P-U are implemented as `torch.nn.Module` object, enabling a simple drop-in replacement of the naive PCC/CCC loss. An example usage is shown below:

```
from b2p import B2PCorrEMA, B2PCorrU

pcc_loss_ema = B2PCorrEMA(metric="PCC", alpha=0.05)
pcc_loss_u = B2PCorrU(metric="CCC", alpha=0.05, reservoir_size=256)

loss = pcc_loss_ema(y_label, y_predict)
loss.backward()
```

B Additional Proofs and Derivations

B.1 Useful identities: U-statistics

Lemma B.1 (U-statistic identities used in §4.2)**.** *Let* (y_1, f_1) *and* (y_2, f_2) *be i.i.d. copies of* $(y, f_\theta(x))$ *with* $\mathbb{E}[y^2] < \infty$ *and* $\mathbb{E}[f^2] < \infty$*. Then*

$$\mathrm{Cov}(y, f) = \frac{1}{2}\,\mathbb{E}\big[(y_1 - y_2)(f_1 - f_2)\big], \tag{14}$$

$$\mathrm{Var}(f) = \frac{1}{2}\,\mathbb{E}\big[(f_1 - f_2)^2\big]. \tag{15}$$

Proof. We expand and use independence. For covariance,

$$\begin{aligned}\mathbb{E}[(y_1 - y_2)(f_1 - f_2)] &= \mathbb{E}[y_1 f_1] - \mathbb{E}[y_1 f_2] - \mathbb{E}[y_2 f_1] + \mathbb{E}[y_2 f_2] \\ &= \mathbb{E}[yf] - \mathbb{E}[y_1]\mathbb{E}[f_2] - \mathbb{E}[y_2]\mathbb{E}[f_1] + \mathbb{E}[yf] \\ &= 2\big(\mathbb{E}[yf] - \mathbb{E}[y]\mathbb{E}[f]\big) \;=\; 2\,\mathrm{Cov}(y, f),\end{aligned}$$

which gives Equation (14). Similarly,

$$\mathbb{E}[(f_1 - f_2)^2] = \mathbb{E}[f_1^2] - 2\mathbb{E}[f_1 f_2] + \mathbb{E}[f_2^2] = 2\mathbb{E}[f^2] - 2(\mathbb{E}[f])^2 = 2\,\mathrm{Var}(f),$$

giving Equation (15). □

B.2 Proof of Lemma 3.1 (population gradient of PCC)

Proof of Lemma 3.1. Assume Assumption 5.1 and write $f = f_\theta(x)$. Recall

$$\rho(\theta) \;=\; \frac{C}{\sigma_y \sigma_f}, \qquad C = \mathbb{E}[(y - \mu_y)(f - \mu_f)], \qquad \sigma_f = \sqrt{\mathrm{Var}(f)}.$$

Since $\mathbb{E}[y - \mu_y] = 0$, we may rewrite

$$C = \mathbb{E}[(y - \mu_y) f].$$

Using differentiation under the expectation,

$$\nabla_\theta C = \nabla_\theta \mathbb{E}[(y - \mu_y) f] = \mathbb{E}[(y - \mu_y)\,\nabla_\theta f]. \tag{16}$$

Next, $\sigma_f = (\sigma_f^2)^{1/2}$ with $\sigma_f^2 = \mathrm{Var}(f) = \mathbb{E}[f^2] - (\mathbb{E}[f])^2$. Again interchanging ∇_θ and $\mathbb{E}$ gives

$$\nabla_\theta \sigma_f^2 = \nabla_\theta \mathbb{E}[f^2] - 2\mu_f \nabla_\theta \mu_f = \mathbb{E}[2f\,\nabla_\theta f] - 2\mu_f\,\mathbb{E}[\nabla_\theta f] = 2\,\mathbb{E}[(f - \mu_f)\nabla_\theta f]. \tag{17}$$

Therefore,

$$\nabla_\theta \sigma_f = \frac{1}{2\sigma_f} \nabla_\theta \sigma_f^2 = \frac{1}{\sigma_f}\,\mathbb{E}[(f - \mu_f)\nabla_\theta f]. \tag{18}$$

Now apply the quotient rule to $\rho = C/(\sigma_y \sigma_f)$, noting σ_y is constant in θ:

$$\begin{aligned}
\nabla_\theta \rho &= \frac{1}{\sigma_y}\left(\frac{\nabla_\theta C}{\sigma_f} - \frac{C}{\sigma_f^2}\nabla_\theta \sigma_f\right) \\
&= \frac{1}{\sigma_y}\left(\frac{1}{\sigma_f}\mathbb{E}[(y-\mu_y)\nabla_\theta f] - \frac{C}{\sigma_f^2}\cdot\frac{1}{\sigma_f}\mathbb{E}[(f-\mu_f)\nabla_\theta f]\right) \\
&= \mathbb{E}\left[\left(\frac{y-\mu_y}{\sigma_y\sigma_f} - \frac{C}{\sigma_y\sigma_f}\frac{f-\mu_f}{\sigma_f^2}\right)\nabla_\theta f\right].
\end{aligned}$$

Since $\rho = C/(\sigma_y\sigma_f)$, we obtain

$$\nabla_\theta \rho = \mathbb{E}\left[\left(\frac{y-\mu_y}{\sigma_y\sigma_f} - \rho\,\frac{f-\mu_f}{\sigma_f^2}\right)\nabla_\theta f\right],$$

which is exactly Theorem 3.1. □

B.3 Proof of Lemma 3.2 (population gradient of CCC)

Proof of Lemma 3.2. Assume Assumption 5.1 and write $f = f_\theta(x)$. Recall

$$\mathrm{CCC}(\theta) = \frac{2C}{D}, \qquad D = \sigma_y^2 + \sigma_f^2 + (\mu_y - \mu_f)^2.$$

We already have $\nabla_\theta C = \mathbb{E}[(y-\mu_y)\nabla_\theta f]$ from Equation (16).

Next compute $\nabla_\theta D$. Since σ_y^2 and μ_y are constants in θ,

$$\nabla_\theta D = \nabla_\theta \sigma_f^2 + \nabla_\theta(\mu_y - \mu_f)^2.$$

From Equation (17), $\nabla_\theta \sigma_f^2 = 2\mathbb{E}[(f-\mu_f)\nabla_\theta f]$. Also, $(\mu_y-\mu_f)^2 = (\mu_f-\mu_y)^2$ so

$$\nabla_\theta(\mu_y-\mu_f)^2 = 2(\mu_f-\mu_y)\,\nabla_\theta \mu_f = 2(\mu_f-\mu_y)\,\mathbb{E}[\nabla_\theta f].$$

Combining and regrouping,

$$\begin{aligned}
\nabla_\theta D &= 2\mathbb{E}[(f-\mu_f)\nabla_\theta f] + 2(\mu_f-\mu_y)\mathbb{E}[\nabla_\theta f] \\
&= 2\mathbb{E}[(f-\mu_y)\nabla_\theta f].
\end{aligned} \tag{19}$$

Finally, apply the quotient rule to $2C/D$:

$$\begin{aligned}
\nabla_\theta \mathrm{CCC} &= \frac{2}{D^2}\Big(D\,\nabla_\theta C - C\,\nabla_\theta D\Big) \\
&= \frac{2}{D^2}\Big(D\,\mathbb{E}[(y-\mu_y)\nabla_\theta f] - C\cdot 2\mathbb{E}[(f-\mu_y)\nabla_\theta f]\Big) \\
&= \mathbb{E}\left[\frac{2}{D^2}\Big(D(y-\mu_y) - 2C(f-\mu_y)\Big)\nabla_\theta f\right],
\end{aligned}$$

which is exactly Theorem 3.2. □

C Proofs for Theoretical Studies

C.1 Pseudo-loss gradients recover the population gradients

Lemma C.1 (Pseudo-losses match population gradients when moments are correct)**.** *Let $\mathcal{L}_\rho$ and $\mathcal{L}_{\mathrm{CCC}}$ be defined in Equations* (4) *and* (5)*. If the population moments $(\mu_y, \sigma_y, \mu_f, \sigma_f, C, D, \rho)$ are used inside* $\mathrm{sg}(\cdot)$*, then*

$$\nabla_\theta \mathbb{E}[\mathcal{L}_\rho(f_\theta(x); y)] = -\nabla_\theta \rho(\theta), \qquad \nabla_\theta \mathbb{E}[\mathcal{L}_{\mathrm{CCC}}(f_\theta(x); y)] = -\nabla_\theta \mathrm{CCC}(\theta).$$

Proof. All moments are wrapped in $\mathrm{sg}(\cdot)$ and thus treated as constants in backpropagation. Hence,

$$\nabla_\theta \mathbb{E}[\mathcal{L}_\rho(f; y)] = \mathbb{E}\left[\frac{\partial \mathcal{L}_\rho}{\partial f} \nabla_\theta f\right].$$

Differentiating Equation (4) w.r.t. f gives

$$\frac{\partial \mathcal{L}_\rho}{\partial f} = -\left(\frac{y - \mu_y}{\sigma_y \sigma_f} - \rho \frac{f - \mu_f}{\sigma_f^2}\right),$$

which matches the negative of the per-example weight in Lemma 3.1. Thus $\nabla_\theta \mathbb{E}[\mathcal{L}_\rho] = -\nabla_\theta \rho$. The CCC case is identical: differentiating Equation (5) yields

$$\frac{\partial \mathcal{L}_{\mathrm{CCC}}}{\partial f} = -\frac{2}{D^2}\Big(D(y - \mu_y) - 2C(f - \mu_y)\Big),$$

which matches the negative weight in Lemma 3.2. □

C.2 Proof of Proposition 5.2 (Bias of naive mini-batch gradients)

Proof of Proposition 5.2 (PCC case). We present the PCC case; CCC follows identically since it is also a smooth ratio of finitely many raw moments of (y, f) on the non-degenerate domain.

Step 1: PCC as a smooth function of raw moments. Fix θ and write $f = f_\theta(x)$. Define

$$r(\theta) := (r_1, r_2, r_3)^\top := \big(\mathbb{E}[f],\ \mathbb{E}[f^2],\ \mathbb{E}[yf]\big)^\top.$$

On the domain $\{r : r_2 - r_1^2 > 0\}$ define

$$\varphi_\rho(r) := \frac{r_3 - \mu_y r_1}{\sigma_y \sqrt{r_2 - r_1^2}}, \qquad \text{so that} \qquad \rho(\theta) = \varphi_\rho(r(\theta)).$$

Given an i.i.d. mini-batch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^B$, let $s_i = (f_i, f_i^2, y_i f_i)^\top$ and $\hat{r} = \frac{1}{B}\sum_{i=1}^B s_i$, so $\widehat{\rho}_\mathcal{B} = \varphi_\rho(\hat{r})$.

Let $J(r) := \nabla_r \varphi_\rho(r)^\top \in \mathbb{R}^{1\times 3}$ and $H(r) := \nabla_r^2 \varphi_\rho(r) \in \mathbb{R}^{3\times 3}$. Under Assumption 5.1(3), φ_ρ is C^3 on a neighborhood of $r(\theta)$ and its derivatives up to third order are bounded there.

Step 2: Taylor expansion and decomposition. Fix any coordinate $k \in \{1, \ldots, d\}$ and write $\partial_k := \partial/\partial\theta_k$. By the chain rule,

$$\partial_k \widehat{\rho}_\mathcal{B} = J(\hat{r})\, \partial_k \hat{r}, \qquad \partial_k \hat{r} = \frac{1}{B}\sum_{i=1}^B \partial_k s_i.$$

Let $\delta r := \hat{r} - r$. A second-order Taylor expansion of J around r yields

$$J(\hat{r}) = J(r) + \delta r^\top H(r) + R, \qquad \text{with} \qquad \|R\| \le C\,\|\delta r\|^2$$

for some constant $C < \infty$ (depending on third derivatives near r). Therefore,

$$\partial_k \widehat{\rho}_\mathcal{B} = \underbrace{J(r)\, \partial_k \hat{r}}_{(i)} + \underbrace{\delta r^\top H(r)\, \partial_k \hat{r}}_{(ii)} + \underbrace{R\, \partial_k \hat{r}}_{(iii)}. \tag{20}$$

Term (i): unbiased leading term. By Assumption 5.1(1), $\mathbb{E}[\partial_k \hat{r}] = \partial_k r$. Hence

$$\mathbb{E}[J(r)\, \partial_k \hat{r}] = J(r)\, \partial_k r = \partial_k \rho.$$

Term (ii): the $O(1/B)$ bias term. Write $u_i := s_i - \mathbb{E}[s] \in \mathbb{R}^3$ and $g_i^{(k)} := \partial_k s_i \in \mathbb{R}^3$. Then

$$\delta r = \frac{1}{B}\sum_{j=1}^B u_j, \qquad \partial_k \hat{r} = \frac{1}{B}\sum_{i=1}^B g_i^{(k)}.$$

Using independence across samples and $\mathbb{E}[u_j] = 0$,

$$\mathbb{E}\big[\delta r\, (\partial_k \hat{r})^\top\big] = \frac{1}{B^2}\sum_{j=1}^B \sum_{i=1}^B \mathbb{E}[u_j (g_i^{(k)})^\top] = \frac{1}{B}\,\mathbb{E}\big[(s - \mathbb{E}[s])(\partial_k s)^\top\big].$$

Since $H(r)$ is bounded on the neighborhood, $\mathbb{E}[\delta r^\top H(r)\, \partial_k \hat{r}] = O(B^{-1})$, and it is generally nonzero because s and $\partial_k s$ come from the same sample.

Term (iii): the Taylor remainder contributes $O(1/B)$. Using $\|R\| \le C\|\delta r\|^2$ and the triangle inequality,

$$\big|\mathbb{E}[R\,\partial_k \hat{r}]\big| \le \mathbb{E}[\|R\|\,\|\partial_k \hat{r}\|] \le C\,\mathbb{E}\big[\|\delta r\|^2\,\|\partial_k \hat{r}\|\big].$$

Let $G^{(k)} := \partial_k \hat{r} = \frac{1}{B}\sum_{i=1}^{B} g_i^{(k)}$. Then

$$\mathbb{E}\big[\|\delta r\|^2\,\|G^{(k)}\|\big] \le \frac{1}{B}\sum_{i=1}^{B}\mathbb{E}\big[\|\delta r\|^2\,\|g_i^{(k)}\|\big] = \mathbb{E}\big[\|\delta r\|^2\,\|g_1^{(k)}\|\big]$$

by exchangeability. Expanding $\|\delta r\|^2$ using $\delta r = \frac{1}{B}\sum_{j=1}^{B} u_j$ gives

$$\|\delta r\|^2 = \frac{1}{B^2}\sum_{j=1}^{B}\sum_{\ell=1}^{B}\langle u_j, u_\ell\rangle.$$

Taking expectation against $\|g_1^{(k)}\|$ and using independence and $\mathbb{E}[u_j] = 0$, all mixed terms vanish, yielding

$$\mathbb{E}\big[\|\delta r\|^2\,\|g_1^{(k)}\|\big] = \frac{B-1}{B^2}\,\mathbb{E}[\|u\|^2]\,\mathbb{E}[\|g^{(k)}\|] \;+\; \frac{1}{B^2}\,\mathbb{E}[\|u\|^2\,\|g^{(k)}\|] = O(B^{-1}),$$

provided the expectations are finite. This holds under Assumption 5.1(2) since by triangle inequality $\|u\|^2 = \|s - \mathbb{E}[s]\|^2 \lesssim 1 + \|s\|^2$ and $\|g^{(k)}\| \le \|\nabla_\theta s\|_F$.

Consequently, $\mathbb{E}[R\,\partial_k \hat{r}] = O(B^{-1})$.

Conclusion. Taking expectations in Equation (20) yields

$$\mathbb{E}_{\mathcal{B}}\big[\partial_k \widehat{\rho}_{\mathcal{B}}\big] = \partial_k \rho + O(B^{-1}).$$

Since this holds for each coordinate k, we obtain $\mathbb{E}_{\mathcal{B}}[\nabla_\theta \widehat{\rho}_{\mathcal{B}}] = \nabla_\theta \rho + O(B^{-1})$. The mini-batch gradient analysis for CCC is identical after defining the corresponding smooth map

$$\varphi_{\mathrm{CCC}}(r) = \frac{2(r_3 - \mu_y r_1)}{\sigma_y^2 + (r_2 - r_1^2) + (\mu_y - r_1)^2} \text{ on the domain } \sigma_y^2 + (r_2 - r_1^2) + (\mu_y - r_1)^2 > 0.$$

□

C.3 Formal Tracking Regime for Theorem 5.3

Assumption C.2 (B2P-EMA tracking regime)**.** Assume Assumption 5.1 holds for all iterates θ_t. Let $m(\theta) := (\mu_f(\theta), \sigma_f^2(\theta), C(\theta))$, $z_t := (\hat{\mu}_f^{(t)}, \widehat{\sigma_f^2}^{(t)}, \hat{C}^{(t)})$, and let $\mathcal{F}_t$ be all information available before drawing the batch at step t. Let $h_t \in \mathbb{R}^3$ be the detached batch-moment estimate used in Equation (8), and let $\bar{g}(\theta, z)$ denote the conditional mean gradient of the B2P-EMA pseudo-loss. Assume:

(1) **Unbiased moment noise:** $\mathbb{E}[h_t \mid \mathcal{F}_t] = m(\theta_t)$ and $\mathbb{E}[\|h_t - m(\theta_t)\|^2 \mid \mathcal{F}_t] \le V$ for some $V < \infty$.

(2) **EMA rates:** $\alpha_t \in (0, 1]$, $\sum_t \alpha_t = \infty$, and $\sum_t \alpha_t^2 < \infty$.

(3) **Slow target drift:** $\|m(\theta_{t+1}) - m(\theta_t)\| \le \alpha_t \beta_t$ a.s. for some $\beta_t \to 0$.

(4) **Stable plug-in gradient:** for the ε-stabilized PCC/CCC pseudo-losses, $\bar{g}(\theta_t, \cdot)$ is Lipschitz, i.e., $\|\bar{g}(\theta_t, z) - \bar{g}(\theta_t, z')\| \le L_g\|z - z'\|$ for some $L_g < \infty$.

C.4 Proof of Theorem 5.3 (Tracking of B2P-EMA)

Proof. Write

$$m_t := m(\theta_t), \qquad z_t := \big(\hat{\mu}_f^{(t)}, \widehat{\sigma_f^2}^{(t)}, \hat{C}^{(t)}\big), \qquad e_t := z_t - m_t.$$

Step 1: Tracking of the moment sketch. The EMA recursion can be written as

$$z_{t+1} = z_t + \alpha_t(h_t - z_t) = z_t + \alpha_t\big(m_t - z_t + \xi_{t+1}\big),$$

where

$$\xi_{t+1} := h_t - m_t.$$

By Assumption C.2(1),

$$\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\xi_{t+1}\|^2 \mid \mathcal{F}_t] \le V.$$

Let

$$\Delta_t := m_{t+1} - m_t.$$

Then

$$e_{t+1} = (1-\alpha_t)e_t + \alpha_t\xi_{t+1} - \Delta_t.$$

Conditioning on $\mathcal{F}_t$, we do not pull Δ_t outside the expectation, since θ_{t+1} may depend on the same batch used to form h_t. Instead, define

$$u_t := (1-\alpha_t)e_t + \alpha_t\xi_{t+1}, \qquad e_{t+1} = u_t - \Delta_t.$$

Using Young's inequality with $\eta_t = \alpha_t/2$,

$$\|e_{t+1}\|^2 \le \Big(1 + \frac{\alpha_t}{2}\Big)\|u_t\|^2 + \Big(1 + \frac{2}{\alpha_t}\Big)\|\Delta_t\|^2.$$

Since e_t is $\mathcal{F}_t$-measurable and $\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = 0$,

$$\begin{aligned}\mathbb{E}[\|u_t\|^2 \mid \mathcal{F}_t] &= (1-\alpha_t)^2\|e_t\|^2 + \alpha_t^2\mathbb{E}[\|\xi_{t+1}\|^2 \mid \mathcal{F}_t] \\ &\le (1-\alpha_t)\|e_t\|^2 + V\alpha_t^2.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[\|e_{t+1}\|^2 \mid \mathcal{F}_t] &\le \Big(1 + \frac{\alpha_t}{2}\Big)\Big((1-\alpha_t)\|e_t\|^2 + V\alpha_t^2\Big) + \Big(1 + \frac{2}{\alpha_t}\Big)\|\Delta_t\|^2 \\ &\le \Big(1 - \frac{\alpha_t}{2}\Big)\|e_t\|^2 + 2V\alpha_t^2 + \Big(1 + \frac{2}{\alpha_t}\Big)\|\Delta_t\|^2.\end{aligned}$$

Applying Assumption C.2(3) and $\alpha_t \in (0, 1]$,

$$\Big(1 + \frac{2}{\alpha_t}\Big)\|\Delta_t\|^2 \le \Big(1 + \frac{2}{\alpha_t}\Big)\alpha_t^2\beta_t^2 \le 3\alpha_t\beta_t^2.$$

Hence

$$\mathbb{E}[\|e_{t+1}\|^2 \mid \mathcal{F}_t] \le \Big(1 - \frac{\alpha_t}{2}\Big)\|e_t\|^2 + 3\alpha_t\beta_t^2 + 2V\alpha_t^2.$$

Taking expectation and writing

$$a_t := \mathbb{E}[\|e_t\|^2]$$

gives

$$a_{t+1} \le \Big(1 - \frac{\alpha_t}{2}\Big)a_t + 3\alpha_t\beta_t^2 + 2V\alpha_t^2.$$

Fix any $\varepsilon > 0$. Since $\beta_t \to 0$, there exists T_ε such that $3\beta_t^2 \le \varepsilon$ for all $t \ge T_\varepsilon$. Hence, for $t \ge T_\varepsilon$,

$$a_{t+1} \le \Big(1 - \frac{\alpha_t}{2}\Big)a_t + \varepsilon\alpha_t + 2V\alpha_t^2.$$

Equivalently, we have

$$a_{t+1} - 2\varepsilon \le \Big(1 - \frac{\alpha_t}{2}\Big)(a_t - 2\varepsilon) + 2V\alpha_t^2.$$

Iterating from T_ε to t yields

$$a_{t+1} - 2\varepsilon \le \prod_{k=T_\varepsilon}^{t}\Big(1 - \frac{\alpha_k}{2}\Big)(a_{T_\varepsilon} - 2\varepsilon) + 2V\sum_{k=T_\varepsilon}^{t}\alpha_k^2\prod_{j=k+1}^{t}\Big(1 - \frac{\alpha_j}{2}\Big).$$

The first term vanishes because $\sum_t \alpha_t = \infty$, and the second vanishes because $\sum_t \alpha_t^2 < \infty$. Therefore

$$\limsup_{t\to\infty} a_t \le 2\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we conclude

$$\mathbb{E}\|z_t - m(\theta_t)\|^2 = a_t \to 0.$$

Step 2: Consistency of the plug-in gradient. Let

$$\bar{g}_t := \mathbb{E}[g_t \mid \mathcal{F}_t] = \bar{g}(\theta_t, z_t).$$

By Lemma C.1, plugging the true moments into the stop-gradient pseudo-loss exactly recovers the population gradient:

$$\bar{g}(\theta_t, m(\theta_t)) = \nabla J(\theta_t).$$

Using Assumption C.2(4) and Jensen's inequality,

$$\begin{aligned}\mathbb{E}\|\bar{g}_t - \nabla J(\theta_t)\| &= \mathbb{E}\|\bar{g}(\theta_t, z_t) - \bar{g}(\theta_t, m(\theta_t))\| \\ &\le L_g\, \mathbb{E}\|z_t - m(\theta_t)\| \\ &\le L_g\, \sqrt{\mathbb{E}\|z_t - m(\theta_t)\|^2} \to 0.\end{aligned}$$

This proves the asymptotic consistency of the moving-target plug-in gradient. □

Remark C.3 (Constant-α heuristic). If $\alpha_t \equiv \alpha \in (0, 1]$ and

$$\sup_t \|m(\theta_{t+1}) - m(\theta_t)\| \le d,$$

then the same one-step recursion gives

$$\limsup_{t\to\infty} \mathbb{E}\|z_t - m(\theta_t)\|^2 \;\le\; \frac{4d^2}{\alpha^2} + 2V\alpha.$$

Thus a small constant EMA still keeps the sketch accurate when the target moments drift on a slower scale than α.

C.5 Proof of Proposition 5.4 (Variance reduction of B2P-U)

Proof of Proposition 5.4. Unbiasedness follows from linearity of expectation:

$$\mathbb{E}[\hat{M}^\star] = \mathbb{E}[\hat{M}_{\mathrm{U}}] + \lambda(\mathbb{E}[\hat{M}_{\mathcal{R}}] - \mathbb{E}[\hat{M}_{\mathcal{R}}]) = M.$$

For the variance,

$$\mathrm{Var}(\hat{M}^\star) = \mathrm{Var}(\hat{M}_{\mathrm{U}}) + \lambda^2 \mathrm{Var}(\hat{M}_{\mathcal{R}}) + 2\lambda\, \mathrm{Cov}(\hat{M}_{\mathrm{U}}, \hat{M}_{\mathcal{R}}).$$

This is a convex quadratic in λ, minimized at $\lambda^\star = -\mathrm{Cov}(\hat{M}_{\mathrm{U}}, \hat{M}_{\mathcal{R}})/\mathrm{Var}(\hat{M}_{\mathcal{R}})$. Substituting $\lambda^\star$ yields the stated variance reduction, with strict inequality when $\mathrm{Cov}(\hat{M}_{\mathrm{U}}, \hat{M}_{\mathcal{R}}) \neq 0$. □

D Detailed Experiment Settings

D.1 Synthetic Dataset

D.1.1 Exp. 1: Imbalanced mixture regression

Exp. 1: Imbalanced mixture regression (synthetic). We construct an imbalanced two-component Gaussian-mixture regression problem to stress-test correlation objectives under mini-batches. Let $d = 64$ and $N_{\mathrm{pop}} = 4000$. We first draw an indicator $z_i \sim \mathrm{Bernoulli}(p_{\mathrm{major}})$ with $p_{\mathrm{major}} = 0.9$, and map it to a sign variable $s_i = 2z_i - 1 \in \{+1, -1\}$, where $s_i = +1$ is the *major* component. Define a mean vector $\mu \in \mathbb{R}^d$ with $\mu_1 = \texttt{mean_sep} = 3.0$ and $\mu_j = 0$ for $j \ge 2$. We then sample features

$$x_i \sim \mathcal{N}(s_i\mu, I_d) \qquad (\text{i.e., } x_i = \xi_i + s_i\mu,\ \xi_i \sim \mathcal{N}(0, I_d)).$$

Next, we sample a ground-truth direction $w^\star \sim \mathcal{N}(0, I_d)$ and normalize it to unit norm. The regression target is generated by

$$y_i = {w^\star}^\top x_i + \Delta\, s_i + \varepsilon_i, \qquad \Delta = \texttt{shift} = 8.0,$$

with heteroscedastic noise

$$\varepsilon_i \sim \mathcal{N}\big(0, \sigma_{s_i}^2\big), \qquad \sigma_{+1} = \texttt{noise_major} = 1.0,\ \ \sigma_{-1} = \texttt{noise_minor} = 4.0.$$

This construction yields well-defined population moments, but makes mini-batch estimates of $\mu_f, \mathrm{Var}(f_\theta), \mathrm{Cov}(y, f_\theta)$ highly variable for small batch size B due to (i) strong class imbalance and (ii) component-dependent label shift and noise.

Exp. 1: Gradient-alignment diagnostics (PCC/CCC). We treat the entire synthesized dataset $\{(x_i, y_i)\}_{i=1}^{N_{\text{pop}}}$ as the *population* to form oracle moments and oracle gradients. For a linear predictor $f_\theta(x) = \theta^\top x$, at each checkpoint $\theta(t)$, we compute an oracle gradient $g_{\text{oracle}}(\theta) = \nabla_\theta \rho(\theta)$ (or $\nabla_\theta \text{ccc}(\theta)$) using the full population. We then estimate stochastic gradients from i.i.d. mini-batches of size B sampled from the population: for each checkpoint we draw $\texttt{n_trials} = 4,096$ batches and compute a per-batch gradient direction $\hat{g} = -\nabla_\theta \ell(\theta; \mathcal{B})$, where ℓ is the naive mini-batch PCC/CCC or our B2P-Corr surrogate. We report:

$$\text{Align} = \cos(\hat{g}, g_{\text{oracle}}), \qquad \text{RelBias} = \frac{\|\mathbb{E}[\hat{g}] - g_{\text{oracle}}\|_2}{\|g_{\text{oracle}}\|_2},$$

where expectations are approximated by averaging over the 4096 sampled mini-batches. We repeat this evaluation for $\texttt{metric} \in \{\text{PCC}, \text{CCC}\}$ and $B \in \{2, 4, 8, 16, 32\}$.

Table 3: Exp. 1 hyperparameters.

Data generation	
Dimension d	64
Population size N_{pop}	4000
Imbalance $p_{\text{major}} = \mathbb{P}(s = +1)$	0.9
Mean separation (`mean_sep`)	3.0 (only on feature 1; others 0)
Label shift Δ (`shift`)	8.0
Noise std (major/minor)	$\sigma_{+1} = 1.0$, $\sigma_{-1} = 4.0$
Data seed	0
Evaluation protocol (gradient diagnostics)	
Metrics	$\{\text{PCC}, \text{CCC}\}$
Batch sizes B	{2, 4, 8, 16, 32}
Checkpoint grid	60 points in $[0, 1]$ (linear interpolation; renormalized)
Mini-batches per checkpoint	$\texttt{n_trials} = 4096$
Random-direction seed (for θ_{rand})	`set_seed(123)` inside each config run
B2P-Corr estimator hyperparameters	
Shared: step size α	0.02
Shared: numerical ϵ	10^{-8}
B2P-U reservoir size	1024

D.1.2 Exp. 2: isotropic linear regression

Exp. 2: Isotropic linear regression (data generation). We construct an isotropic linear regression problem in $\mathbb{R}^d$ with additive Gaussian noise. Fix the dimension $d = 256$ and sample the ground-truth direction $w^\star \sim \mathcal{N}(0, I_d)$, then normalize it to unit norm $\|w^\star\|_2 = 1$. Next, we generate i.i.d. covariates $x_i \sim \mathcal{N}(0, I_d)$ and responses

$$y_i = {w^\star}^\top x_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),\ \sigma = 3.0, \tag{21}$$

for a total of $N_{\text{train}} + N_{\text{test}} = 50{,}000 + 10{,}000$ samples. We split the first N_{train} examples for training and the remaining N_{test} for testing.

Exp. 2: Mini-batch PCC optimization. We train a bias-free linear predictor $f_\theta(x) = \theta^\top x$ (implemented as `Linear(d,1,bias=False)`), where θ is initialized as $\theta \sim \mathcal{N}(0, 10^{-4} I_d)$ (std. 10^{-2}). We optimize the PCC objective using Adam for 5,000 steps with mini-batch size B (we run separate experiments for each B). We evaluate (i) *DirectionRecovery* $= \cos(\theta, w^\star)$ and (ii) *TestPCC* computed on the *full* test set. Metrics are logged every 100 steps (and at step 1) and averaged over 3 random seeds (which control model initialization and data-loader shuffling).

Table 4: Exp. 2 hyperparameters

Data generation	
Dimension d	256
Train / test size	$N_{\text{train}} = 50{,}000$, $N_{\text{test}} = 10{,}000$
Covariates	$x \sim \mathcal{N}(0, I_d)$
Ground-truth direction	$w^\star \sim \mathcal{N}(0, I_d)$, normalized to $\|w^\star\|_2 = 1$
Response model	$y = {w^\star}^\top x + \varepsilon$
Noise	$\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 3.0$
Dataset seed	0
Training & evaluation protocol	
Initialization	$\theta \sim \mathcal{N}(0, 10^{-4} I)$ (`std` $=10^{-2}$)
Batch size B	$B \in \{2, 4, 8, 16\}$
Training steps	$5,000$
Optimizer	Adam
Learning rate	5×10^{-3}
Weight decay	10^{-3}
Gradient clipping	$\|\nabla\|_2 \leq 10$
B2P-Corr estimator hyperparameters	
Shared: step size α	0.05
Shared: numerical ϵ	10^{-8}
B2P-U reservoir size	256

D.2 Real-world Dataset

D.2.1 Video quality assessment (KoNViD-1k)

The hyperparameter settings for the KoNViD-1k experiment are provided in Table 5. We strictly follow the data processing and training pipeline of the baseline KVQ [16], use the same reported metrics (PCC, SROCC, and RMSE), and only replace the naive mini-batch PCC objective with our decomposable B2P-Corr and B2P-U objectives.

Table 5: KVQ [16] experiment setting on KoNViD-1k dataset

Model & input	
Clip length	32
Input channels	768
Hidden channels	64
Optimization schedule	
Epochs	60
Warmup epochs	1
Batch size	$B \in \{2, 4, 8\}$
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	0.05
B2P-Corr estimator hyperparameters	
Shared: step size α	0.05
Shared: numerical ϵ	10^{-8}
B2P-U reservoir size	256

D.2.2 Spatial transcriptomic (10xProteomic)

The hyperparameter settings for the 10xProteomic experiment are provided in Table 6. We strictly follow the data processing and training pipeline of the baseline EGN [25], use the same reported

metrics (PCC, MAE and MSE), and only replace the naive mini-batch PCC objective with our decomposable B2P-Corr and B2P-U objectives.

Table 6: EGN [25] experiment setting on 10xProteomic dataset

Model & input	
Number of genes	250
MLP dimension	4096
Hidden dimension	1024
Model depth	8
Optimization schedule	
Epochs	50
Warmup steps	100
Batch size	$B \in \{8, 16, 32\}$
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	1×10^{-4}
B2P-Corr estimator hyperparameters	
Shared: step size α	0.05
Shared: numerical ϵ	10^{-8}
B2P-U reservoir size	256

D.2.3 Video-based sentiment analysis (VEATIC)

The hyperparameter settings for the VEATIC experiment are provided in Table 7. We strictly follow the data processing and training pipeline in VEATIC [17], use the same reported metrics (CCC, PCC, RMSE), and only replace the naive mini-batch CCC objective with our decomposable B2P-Corr and B2P-U objectives.

Table 7: VEATIC [17] experiment setting

Model & input	
MLP dimension	2048
Hidden dimension	2048
Model depth	6
Dropout	0.4
Backbone	resnet50
Optimization schedule	
Epochs	20
Batch size	$B \in \{4, 10, 20\}$
Optimizer	Adam
Learning rate	5×10^{-4}
Weight decay	1×10^{-8}
Betas	$[0.9, 0.999]$
B2P-Corr estimator hyperparameters	
Shared: step size α	0.05
Shared: numerical ϵ	10^{-8}
B2P-U reservoir size	256

D.3 Hardware Resources

All experiments were conducted on a server running Ubuntu 22.04, equipped with 6 NVIDIA A100 GPUs (80GB VRAM each) and 512GB of system RAM.

E Extended Results and Analysis

E.1 VEATIC Dataset Arousal Result

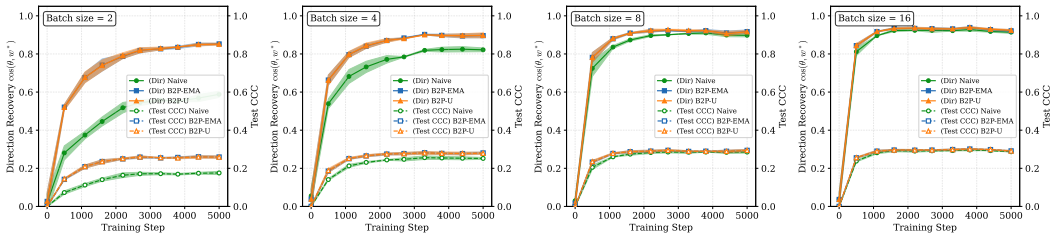
Table 8: **VEATIC arousal (CCC) results.** We keep the baseline architecture, optimizer and auxiliary losses fixed and replace only the naive mini-batch CCC objective with B2P-Corr (B2P-EMA/B2P-U). B2P consistently improves CCC/PCC and reduces RMSE across batch sizes. "Default" denotes the batch size used in the original baseline. Results are reported as mean±std (bold/underline denote best/second best).

Dataset	Batch-size	4			10			20 (Default)		
	Metrics	CCC ↑	PCC ↑	RMSE ↓	CCC ↑	PCC ↑	RMSE ↓	CCC ↑	PCC ↑	RMSE ↓
VEATIC Arousal (CCC)	VEATIC [17]	$0.433_{\pm 0.025}$	$0.515_{\pm 0.019}$	$0.311_{\pm 0.017}$	$0.560_{\pm 0.033}$	$0.602_{\pm 0.029}$	$0.230_{\pm 0.012}$	$0.577_{\pm 0.024}$	$0.619_{\pm 0.030}$	$0.234_{\pm 0.015}$
	+B2P-EMA	$\mathbf{0.452}_{\pm 0.023}$	$\underline{0.538}_{\pm 0.031}$	$\underline{0.246}_{\pm 0.015}$	$\underline{0.569}_{\pm 0.036}$	$\underline{0.612}_{\pm 0.021}$	$\underline{0.216}_{\pm 0.017}$	$\underline{0.591}_{\pm 0.028}$	$\underline{0.630}_{\pm 0.020}$	$\underline{0.217}_{\pm 0.014}$
	+B2P-U	$\underline{0.450}_{\pm 0.027}$	$\mathbf{0.541}_{\pm 0.019}$	$\mathbf{0.218}_{\pm 0.014}$	$\mathbf{0.578}_{\pm 0.025}$	$\mathbf{0.630}_{\pm 0.033}$	$\mathbf{0.210}_{\pm 0.009}$	$\mathbf{0.603}_{\pm 0.019}$	$\mathbf{0.643}_{\pm 0.023}$	$\mathbf{0.212}_{\pm 0.010}$

Table 8 reports arousal prediction on VEATIC. Keeping the baseline setting unchanged and replacing only the naive mini-batch CCC term with B2P-Corr yields consistent gains across batch sizes. At B=4, B2P-U improves CCC from 0.433 to 0.450 and sharply reduces RMSE from 0.311 to 0.218 (B2P-EMA: CCC 0.452, RMSE 0.246). The improvements persist at B=10 (best CCC 0.578 vs. 0.560, RMSE 0.210 vs. 0.230) and at the default B=20, where B2P-U achieves the best overall CCC/PCC (0.603/0.643) with lower RMSE (0.212 vs. 0.234). Overall, B2P provides a practical and effective drop-in replacement for stabilizing CCC optimization on this sentimental video-based analysis dataset.

E.2 Direction recovery and test set CCC for synthetic Exp. 2

Figure 4 reports DIRECTIONRECOVERY and CCC on the isotropic linear-regression test set for batch sizes $B \in \{2, 4, 8, 16\}$. The overall behavior closely mirrors the PCC results in Figure 3. In the extreme mini-batch regime ($B = 2$), both B2P-EMA and B2P-U remain stable and continue to improve throughout training, reaching ≈ 0.85 DIRECTIONRECOVERY and ≈ 0.27 CCC. Notably, this CCC is comparable to the naive method trained with a much larger batch size ($B = 16$, ≈ 0.29). In contrast, naive mini-batch CCC optimization degrades sharply at $B = 2$, attaining only ≈ 0.18 CCC and exhibiting substantially worse alignment with the ground-truth direction. Overall, Exp. 2 demonstrates that B2P-Corr enables reliable CCC maximization in the micro-batch setting, where the naive mini-batch objective suffers from severe batch-to-population bias and can become unstable.



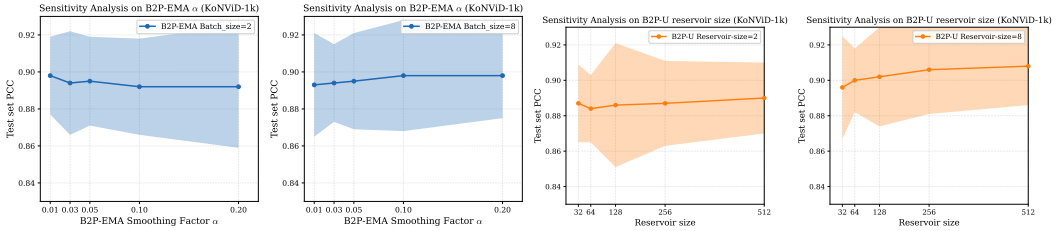
(a) CCC: training ($B = 2$) (b) CCC: training ($B = 4$) (c) CCC: training ($B = 8$) (d) CCC: training ($B = 16$)

Figure 4: For Exp. 2, we show DirectionRecovery and TestCCC vs. training steps for different batch sizes. With B=2, B2P-EMA/B2P-U shows largest increment comparing to naive mini-batch optimization. The gap narrows as B increases.

E.3 B2P-EMA hyperparameter α sensitivity

Figures 5a and 5b report test-set PCC on KoNViD-1k for B2P-EMA with different smoothing factors α under B=2 and B=8. Overall, B2P-EMA is stable across hyperparameter choices. With small batches (B=2), smaller smoothing factors perform better: a smaller α updates the running moment sketch less aggressively, which is desirable because moment estimates from very small batches are noisier. Consequently, α=0.01 outperforms α=0.2 (0.898 vs. 0.893). When increasing the batch size

to B=8, moment estimates concentrate, and larger α values (e.g., α=0.10 or α=0.20) outperform α=0.01 (0.898 vs. 0.893). Overall, we recommend using a larger smoothing factor for larger batch sizes.



(a) B2P-EMA α ($B = 2$) (b) B2P-EMA α ($B = 8$) (c) B2P-U Reservoir (B =2) (d) B2P-U Reservoir (B = 8)

Figure 5: Sensitivity analysis on the KoNViD-1k dataset. We replace the naive mini-batch PCC objective with B2P-Corr and B2P-U, and evaluate (a–b) test-set PCC of B2P-EMA under different smoothing factors α, and (c–d) test-set PCC of B2P-U under different reservoir sizes, for batch sizes $B \in \{2, 8\}$.

E.4 B2P-U reservoir size sensitivity

Figures 5c and 5d report test-set PCC on KoNViD-1k for B2P-U with different reservoir sizes under B=2 and B=8. B2P-U remains stable once the reservoir size exceeds 32. For both B=2 and B=8, test-set PCC generally increases with reservoir size. However, larger reservoirs also increase the risk of stale cached predictions (if the reservoir is refreshed infrequently) and add computational overhead.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims regarding theoretical guarantees (bias quantification, tracking, variance reduction) and empirical improvements, which are fully supported by the proofs in Section 5, Appendices B-C and the experimental results in Section 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Theoretical limitations and required conditions are explicitly stated in Assumption 5.1 and Assumption C.2. Practical limitations, such as potential bias introduced by stale reservoir predictions (Remark 5.5) and the computational overhead of larger reservoirs (Section 6.3), are also discussed.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The full set of theoretical assumptions is explicitly detailed in Section 5 (Assumption 1) and Appendix C.3 (Assumption 2). Complete formal proofs for all lemmas, propositions, and theorems are provided in Appendix B and C.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Comprehensive details required for reproducibility, including data generation, model architectures, data splits, evaluation metrics, and optimization schedules for both synthetic and real-world benchmarks, are documented in Section 6 and Appendix D. We also provide the code for B2P-EMA implementation and the synthetic Exp.1 in supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: As stated in Appendix A, an anonymized PyTorch implementation of the B2P-Corr modules and the code to reproduce the synthetic experiments are provided in the supplementary material, with a commitment to release the full reproduction codebase upon acceptance. Moreover, all the real-world datasets we use are open access.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed experimental configurations, including data splits, network architectures, optimizers, learning rates, batch sizes, and specific hyperparameters for all experiments, are explicitly specified in Section 6 and Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The experimental results for the real-world benchmarks report both the mean and standard deviation (error bars indicated with $\pm$ subscripts) across multiple random seeds or cross-validation folds (e.g., 3-fold for Spatial Transcriptomic dataset), as shown in Table 1 and Appendix E.1. Synthetic results are averaged over thousands of trials (e.g., n_trials=4096).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We explicitly detail the hardware compute resources (NVIDIA A100 80GB GPUs, 512GB RAM) used for our experiments in the Appendix, and we quantitatively report the per-iteration runtime (Sec/iter) and peak GPU memory consumption (VRAM) in Section 6.3 (Table 2).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research introduces foundational mathematical optimization algorithms and evaluates them on standard, publicly available datasets, fully conforming with the NeurIPS Code of Ethics.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper focuses entirely on foundational machine learning methodology (stochastic optimization for non-decomposable correlation objectives). As a general algorithmic contribution, it is not connected to specific applications that pose direct, foreseeable negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper proposes a general mathematical optimization method and does not release models or datasets with a high risk of malicious misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: We properly credit and cite the original creators of all existing datasets (KoNViD-1k, 10xProteomic, VEATIC) and baseline codebases used in our experiments. However, due to space constraints, we did not explicitly list their specific license names in the context.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The newly introduced assets (the B2P-Corr codebase and experimental scripts) are documented with a usage example in Appendix A, and the anonymized code is included in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The research relies entirely on existing, publicly available datasets and synthetic data, and does not involve any crowdsourcing or new primary data collection from human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve human subjects, making IRB approval inapplicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large Language Models (LLMs) were not used as an important, original or non-standard component of the core methodology, mathematical proofs, or scientific research in this paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.