

362 A SMISO

363 In this section, we will have a brief introduction to SMISO [1]. Assume we have a loss function of
364 the form

$$\mathbb{E}_{n, \epsilon} f(w; n, \epsilon) \quad (17)$$

365 Similar to SAGA [6], SMISO maintains a parameter table $W = \{w^1, \dots, w^N\}$ which stores the
366 parameter value the last time each data point was accessed. SMISO then maintains an average of
367 the value in the parameter table $\bar{w}_k = \mathbb{E}_n w_k^n$ where k denotes the k_{th} iteration. \bar{w}_k will later be used
368 as the point for gradient evaluation. Given a randomly drawn sample n and ϵ , SMISO would first
369 update the n_{th} entity in W using exponential average

$$w_n^{k+1} = (1 - \alpha)w_n^k + \alpha(\bar{w}_k - \gamma \nabla f(\bar{w}_k; \epsilon, n)). \quad (18)$$

370 Then, it updates \bar{w}_k using running average

$$\bar{w}_{k+1} = \bar{w}_k + \frac{1}{N}w_n^{k+1} - \frac{1}{N}w_n^k. \quad (19)$$

371 If we expand the equation above, we get

$$\bar{w}_{k+1} = \bar{w}_k + \frac{1}{N}w_n^{k+1} - \frac{1}{N}w_n^k \quad (20)$$

$$= \bar{w}_k + \frac{1}{N}[(1 - \alpha)w_n^k + \alpha(\bar{w}_k - \gamma \nabla f(\bar{w}_k; \epsilon, n)) - w_n^k] \quad (21)$$

$$= \bar{w}_k - \frac{\alpha}{N}[\gamma \nabla f(\bar{w}_k; \epsilon, n) + w_n^k - \bar{w}_k] \quad (22)$$

$$= \bar{w}_k - \frac{\alpha}{N}[\gamma \nabla f(\bar{w}_k; \epsilon, n) - (\bar{w}_k - w_n^k)] \quad (23)$$

372 In this case, $\alpha\gamma/N$ is the effective step size. Notice that, if we are using a mini-batch of in-
373 dices/samples, denoted as $B = \{n_b\}$, in which case multiple entities in the parameter table would be
374 updated in an iteration, then we would have

$$\bar{w}_{k+1} = \bar{w}_k + \sum_{n_b \in B} \left[\frac{1}{N}w_{n_b}^{k+1} - \frac{1}{N}w_{n_b}^k \right] \quad (24)$$

$$= \bar{w}_k - \frac{\alpha|B|}{N} \left[\gamma \mathbb{E}_{n_b} \nabla f(\bar{w}_k; \epsilon, n_b) - \mathbb{E}_{n_b} (\bar{w}_k - w_{n_b}^k) \right] \quad (25)$$

375 in which case the effective step size would become $\frac{\alpha|B|\gamma}{N}$. Therefore, in order to compare SMISO
376 with other estimators using SGD under the same step size, we can first select a range of step sizes for
377 SMISO $\{\gamma_0, \gamma_1, \dots\}$ and test SGD with step sizes of

$$\left\{ \frac{\alpha|B|\gamma}{N}\gamma_0, \frac{\alpha|B|\gamma}{N}\gamma_1, \dots \right\}. \quad (26)$$

378 It is also worth mentioning that, it is not clear to us how to introduce momentum or adaptive step
379 size into SMISO, as we have to strictly follow the running mean update formula (Eq. (19)) to ensure
380 $\mathbb{E}_n(\bar{w}_k - w_n^k) = 0$ for unbiasedness. Adding additional terms (e.g. momentum) or changing the
381 scale of the updates (e.g. normalizing the update by its norm) without careful design could break the
382 unbiasedness. However, studying such modifications is beyond the scope of our paper therefore we
383 only compare our methods with SMISO in its original form.

384 B Derivation of variance for different estimators

385 In this section, we will show the full derivation for the trace of the variance of g_{cv} , g_{inc} and g_{combo} .

386 B.1 Variance of g_{cv}

387 In this section, we will derive the trace for the cv estimator defined as

$$g_{cv}(w; n, \epsilon) = \nabla f(w; n, \epsilon) + \underbrace{\mathbb{E}_{\eta} \nabla \tilde{f}(w; n, \eta) - \nabla \tilde{f}(w; n, \epsilon)}_{c_{cv}(w; n, \epsilon)}, \quad (27)$$

388 where \tilde{f} is an approximation function of f with closed-form expectation with respect to ϵ .

389 To start with, we will apply the law of total variance

$$\mathbb{V}[g_{cv}] = \mathbb{E}_{\epsilon} \mathbb{V}_{\eta} g_{cv} + \mathbb{V}_{\epsilon} \mathbb{E}_{\eta} g_{cv}. \quad (28)$$

390 The first term can be computed as

$$\mathbb{E}_{\epsilon} \mathbb{V}_{\eta} g_{cv} = \mathbb{E}_{\epsilon} \mathbb{V}_{\eta} [\nabla f(w; n, \epsilon) + \mathbb{E}_{\eta} \nabla \tilde{f}(w; n, \eta) - \nabla \tilde{f}(w; n, \epsilon)] \quad (29)$$

$$= \mathbb{E}_{\epsilon} \mathbb{V}_{\eta} [\nabla f(w; n, \epsilon) - \nabla \tilde{f}(w; n, \epsilon)], \quad (30)$$

391 which follows since $\mathbb{E}_{\eta} \nabla \tilde{f}(w; n, \eta)$ is a constant with respect to ϵ and therefore does not affect the
392 variance.

393 The second term can be computed as

$$\mathbb{V}_{\epsilon} \mathbb{E}_{\eta} g_{cv} = \mathbb{V}_{\epsilon} \mathbb{E}_{\eta} [\nabla f(w; n, \epsilon) + \mathbb{E}_{\eta} \nabla \tilde{f}(w; n, \eta) - \nabla \tilde{f}(w; n, \epsilon)] \quad (31)$$

$$= \mathbb{V}_{\epsilon} \left[\mathbb{E}_{\epsilon} [\nabla f(w; n, \epsilon)] + \mathbb{E}_{\epsilon} [\mathbb{E}_{\eta} \nabla \tilde{f}(w; n, \eta)] - \mathbb{E}_{\epsilon} [\nabla \tilde{f}(w; n, \epsilon)] \right] \quad (32)$$

$$= \mathbb{V}_{\epsilon} \mathbb{E}_{\eta} [\nabla f(w; n, \epsilon) + \nabla \tilde{f}(w; n) - \nabla \tilde{f}(w; n)] \quad (33)$$

$$= \mathbb{V}_{\epsilon} \mathbb{E}_{\eta} [\nabla f(w; n, \epsilon)] \quad (34)$$

$$= \mathbb{V}_{\epsilon} [\nabla f(w; n)]. \quad (35)$$

394 Then we can combine the two terms together to get

$$\mathbb{V}[g_{cv}] = \mathbb{E}_{\epsilon} \mathbb{V}_{\eta} [\nabla f(w; n, \epsilon) - \nabla \tilde{f}(w; n, \epsilon)] + \mathbb{V}_{\epsilon} [\nabla f(w; n)] \quad (36)$$

395 B.2 Variance of g_{inc}

396 Here, we will derive the trace of the variance of the inc estimator defined as

$$g_{inc}(w; n, \epsilon) = \nabla f_n(w; n, \epsilon) + \underbrace{\mathbb{E}_{\mathbf{m}} \nabla f(w^{\mathbf{m}}; \mathbf{m}, \epsilon) - \nabla f(w^n; n, \epsilon)}_{c_{inc}(w; n, \epsilon)}. \quad (37)$$

397 We can derive its variance by first applying the law of total variance

$$\mathbb{V}[g_{inc}] = \mathbb{E}_{\epsilon} \mathbb{V}_{\mathbf{n}} g_{inc} + \mathbb{V}_{\epsilon} \mathbb{E}_{\mathbf{n}} g_{inc}. \quad (38)$$

398 The first term can be computed as

$$\mathbb{E}_{\epsilon} \mathbb{V}_{\mathbf{n}} g_{inc} = \mathbb{E}_{\epsilon} \mathbb{V}_{\mathbf{n}} [\nabla f_n(w; n, \epsilon) + \mathbb{E}_{\mathbf{m}} \nabla f(w^{\mathbf{m}}; \mathbf{m}, \epsilon) - \nabla f(w^n; n, \epsilon)] \quad (39)$$

$$= \mathbb{E}_{\epsilon} \mathbb{V}_{\mathbf{n}} [\nabla f(w; n, \epsilon) - \nabla f(w^n; n, \epsilon)], \quad (40)$$

399 where the second line follows because $\mathbb{E}_{\mathbf{m}} \nabla f(w^{\mathbf{m}}; \mathbf{m}, \epsilon)$ is a constant with respect to n .

400 The second term can be computed as

$$\mathbb{V}_{\epsilon, n} g_{\text{inc}} = \mathbb{V}_{\epsilon, n} [\nabla f_n(w; n, \epsilon) + \mathbb{E}_m \nabla f(w^m; m, \epsilon) - \nabla f(w^n; n, \epsilon)] \quad (41)$$

$$= \mathbb{V}_{\epsilon} \left[\mathbb{E}_n \nabla f_n(w; n, \epsilon) + \mathbb{E}_m \mathbb{E}_n \nabla f(w^m; m, \epsilon) - \mathbb{E}_n \nabla f(w^n; n, \epsilon) \right] \quad (42)$$

$$= \mathbb{V}_{\epsilon} \left[\mathbb{E}_n [\nabla f_n(w; n, \epsilon)] + \mathbb{E}_m \nabla f(w^m; m, \epsilon) - \mathbb{E}_n \nabla f(w^n; n, \epsilon) \right] \quad (43)$$

$$= \mathbb{V}_{\epsilon, n} [\nabla f_n(w; n, \epsilon)] \quad (44)$$

$$= \mathbb{V}_{\epsilon} [\nabla f(w; \epsilon)], \quad (45)$$

401 which then leads us to

$$\mathbb{V}[g_{\text{inc}}] = \mathbb{E}_{\epsilon, n} [\nabla f(w; n, \epsilon) - \nabla f(w^n; n, \epsilon)] + \mathbb{V}_{\epsilon} [\nabla f(w; \epsilon)]. \quad (46)$$

402 B.3 Variance of g_{combo}

403 In this section, we will derive the variance for the estimator g_{combo} defined as

$$g_{\text{combo}}(w; n, \epsilon) = \nabla f(w; n, \epsilon) + \underbrace{\beta c_{\text{cv}}(w; n, \epsilon) + (1 - \beta) c_{\text{inc}}(w; n, \epsilon)}_{c_{\text{combo}}(w; n, \epsilon)}, \quad (47)$$

404 under the ideal assumption where we have $f = \tilde{f}$ and $w = w^n, \forall n$. The variance can be derived
405 through

$$\mathbb{V}[g_{\text{combo}}] = \mathbb{V}_{\epsilon, n} [\nabla f(w; n, \epsilon) + \beta c_{\text{cv}}(n, \epsilon) + (1 - \beta) c_{\text{inc}}(n, \epsilon)] \quad (48)$$

$$= \mathbb{V}_{\epsilon, n} \left[\nabla f(w; n, \epsilon) + \beta \left(\mathbb{E}_{\eta} \nabla \tilde{f}(w; n, \eta) - \nabla \tilde{f}(w; n, \epsilon) \right) + \right. \\ \left. (1 - \beta) \left(\nabla \mathbb{E}_m f(w^m; m, \epsilon) - \nabla f(w^n; n, \epsilon) \right) \right] \quad (49)$$

406 Then we replace \tilde{f} with f and w^n with w based on our assumption,

$$\mathbb{V}[g_{\text{combo}}] = \mathbb{V}_{\epsilon, n} \left[\nabla f(w; n, \epsilon) + \beta \left(\mathbb{E}_{\eta} \nabla f(w; n, \eta) - \nabla f(w; n, \epsilon) \right) + \right. \\ \left. (1 - \beta) \left(\nabla \mathbb{E}_m f(w; m, \epsilon) - \nabla f(w; n, \epsilon) \right) \right] \quad (50)$$

$$= \mathbb{V}_{\epsilon, n} \left[\nabla f(w; n, \epsilon) + \beta (\nabla f(w; n) - \nabla f(w; n, \epsilon)) + (1 - \beta) (f(w; \epsilon) - f(w; n, \epsilon)) \right] \quad (51)$$

$$= \mathbb{V}_{\epsilon, n} \left[\beta \nabla f(w; n) + (1 - \beta) \nabla f(w; \epsilon) \right] \quad (52)$$

$$= \beta^2 \mathbb{V}_n [\nabla f(w; n)] + (1 - \beta)^2 \mathbb{V}_{\epsilon} [\nabla f(w; \epsilon)]. \quad (53)$$

407 The last line follows because $\nabla f(w; n)$ is independent of $\nabla f(w; \epsilon)$.

408 C Step-size search range

For Australian and Sonar, we experiment with learning rates of

$$\{7.5\text{e-}3, 5\text{e-}3, 2.5\text{e-}3, 1\text{e-}3, 5\text{e-}4, 1\text{e-}4, 5\text{e-}5, 2.5\text{e-}5, 1\text{e-}5\}$$

For MNIST, PPCA and Tennis, we used

$$\{1\text{e-}1, 5\text{e-}2, 1\text{e-}2, 5\text{e-}3, 1\text{e-}3\}$$

409 for naive, cv and dual, where the optimizer is Adam.

When optimizing with SMISO, we set $\alpha = 0.9$ and we perform grid search over the value of γ , for MNIST with SMISO, we experiment with γ in

$$\{5e-2, 2.5e-2, 1e-2, 5e-3, 2.5e-3, 1e-3, 5e-4, 1e-4, 5e-5, 1e-5\}$$

For Tennis with SMISO, we experiment with γ in

$$\{5e-2, 2.5e-2, 1e-2, 5e-3, 1e-3, 1e-4, 1e-5\}$$

For PPCA with SMISO, we experiment with γ in

$$\{1e-2, 5e-3, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7\}$$

410 D Generic optimization algorithm

411 In Alg. 1, we describe the end-to-end procedure of applying dual control variate in BBVI. The dual
412 control variate can also be applied in generic doubly-stochastic optimization problems as is shown in
413 Alg. 2.

Algorithm 2 Dual control variate for generic doubly-stochastic optimization problem

Require: Learning rate λ , doubly-stochastic objective $f(w; n, \epsilon)$, approximation $\tilde{f}(w; n, \epsilon)$ where $\mathbb{E}_\epsilon \tilde{f}(w; n, \epsilon)$ is tractable.

Initialize the parameter w_0 , the parameter table $W = \{w^1, \dots, w^N\}$ and the running mean $M = \mathbb{E}_m \mathbb{E}_\eta \nabla \tilde{f}(w_0; m, \eta)$.

for $k = 1, 2, \dots$ **do**

 Sample n and ϵ .

 Extract the value of w^n from the table W .

 Compute the base gradient $g \leftarrow \nabla f(w_k; n, \epsilon)$.

 Compute the control variate $c \leftarrow M - \nabla \tilde{f}(w^n; n, \epsilon)$. ▷ Uses that $M = \mathbb{E}_m \mathbb{E}_\eta \nabla \tilde{f}(w^m; m, \eta)$.

 Update the running mean $M \leftarrow M + \frac{1}{N} (\mathbb{E}_\eta \nabla \tilde{f}(w_k; n, \eta) - \mathbb{E}_\eta \nabla \tilde{f}(w^n; n, \eta))$

 Update the table $w^n \leftarrow w_k$ and update the parameter $w_{k+1} \leftarrow w_k - \lambda(g + c)$.

▷ Or use $g + c$ as a gradient estimator in any stochastic optimization algorithm.

end for

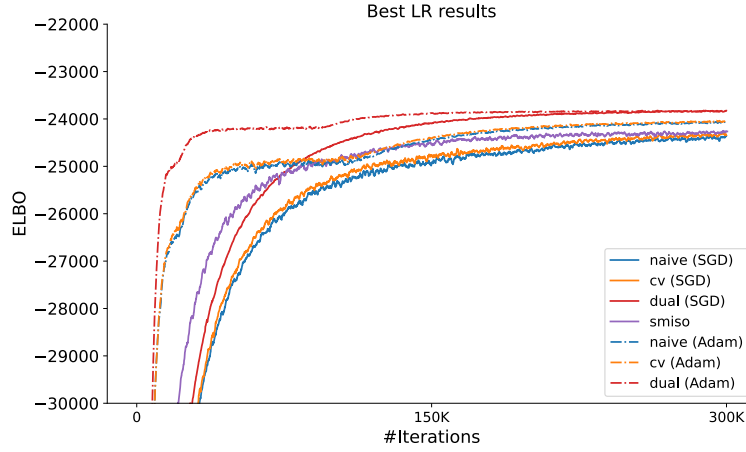
414 E Additional experiment results

415 In this section, we compare naive, cv, and dual with SMISO using SGD. The step sizes for SMISO
416 are the same as the values shown in Sec. C. The step sizes for SGD are converted through Eq. (26)
417 correspondingly. Additionally, we compare their performance with the optimization results acquired
418 using Adam. The results are presented in Fig. 5 and Fig. 6. Overall, we notice that, with SGD, dual
419 still shows superior performance compared with SMISO. In addition, all estimators show performance
420 slower than that of Adam when optimized with SGD except for dual on Tennis.

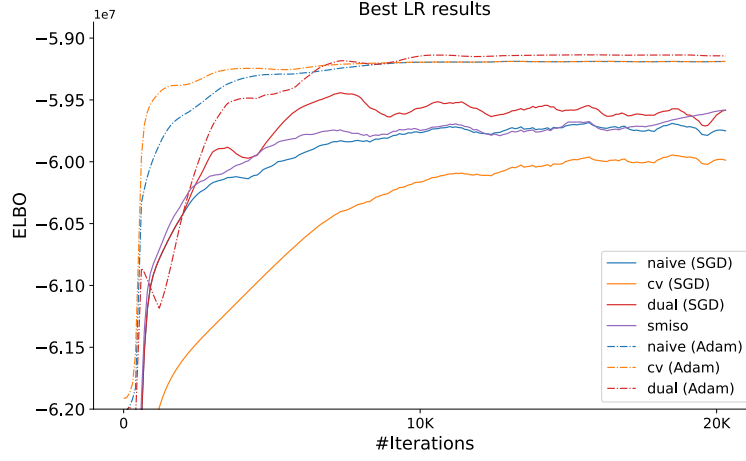
421 Note that, when experimenting with PPCA using dual and SGD, we perform updates with naive
422 in the first three epochs to avoid diverging, as the dual shows a high gradient norm in the first few
423 epochs when SAGA is still warming up. This modification is not required when using Adam, as
424 Adam adaptively chooses the step size based on the gradient norm.

425 F Wall clock time v.s convergence

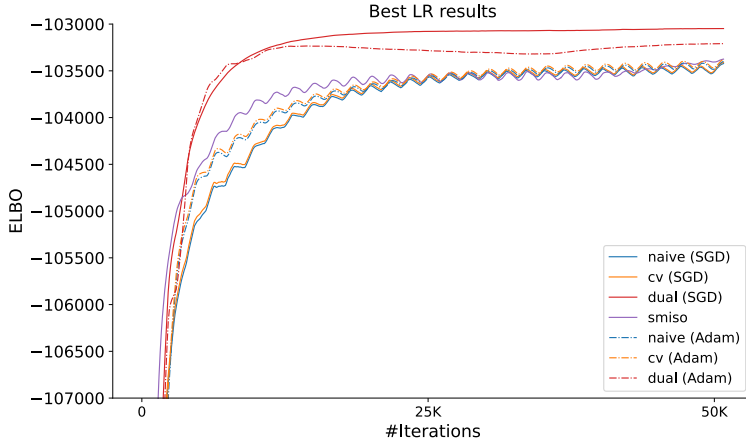
426 In this section, we provide the wall clock time v.s. convergence results. The results are presented in
427 Fig. 7. The results are identical to the results in the second column in Fig. 4 with the x-axis for each
428 estimator rescaled using the values from Table. 2.



(a) MNIST



(b) PPCA



(c) Tennis

Figure 5: **Comparison of different estimators on MNIST, PPCA, and Tennis under SGD and Adam.** The proposed dual combined with Adam shows the best performance on all tasks except Tennis, in which dual with SGD demonstrates the best convergence. For other estimators, Adam leads to better and faster convergence than SGD.

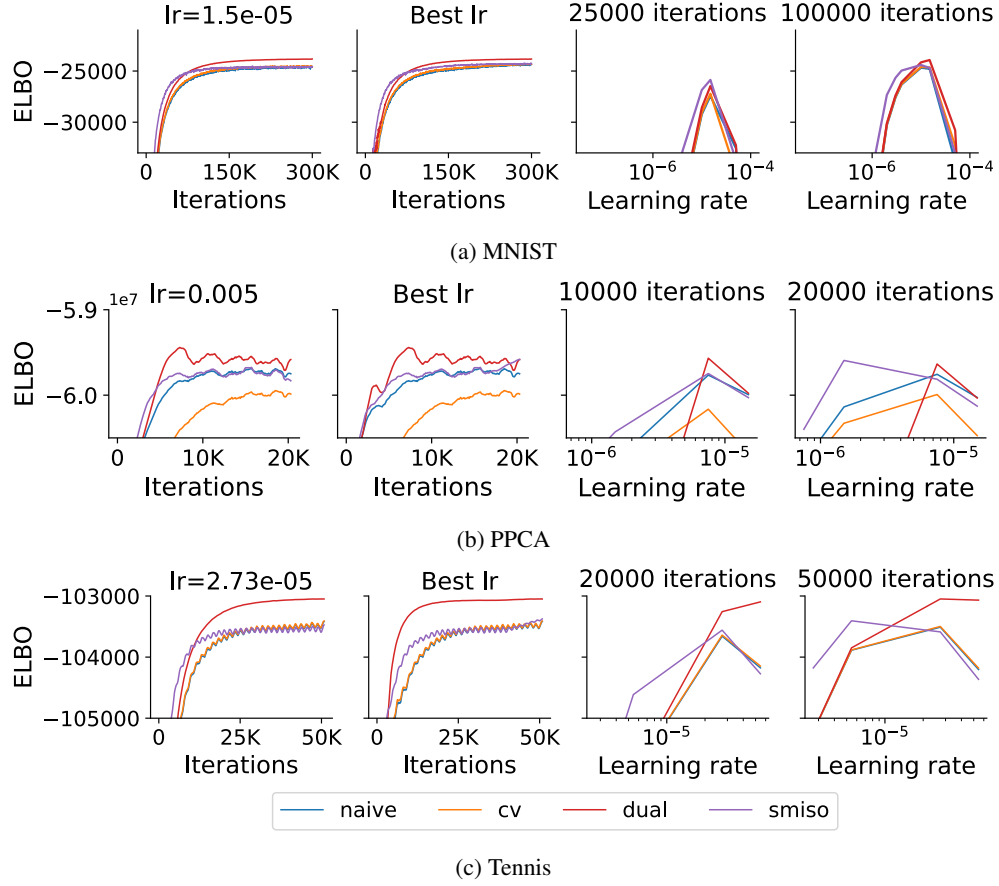


Figure 6: **Optimization results on MNIST, PPCA, and Tennis with SGD.** Using SGD does not affect the improvement of dual against naive and cv. In addition, we notice that dual still performs better than SMISO under SGD, we suspect that this is because dual marginalizes ϵ out explicitly while SMISO approximates the expectation using exponential averaging.

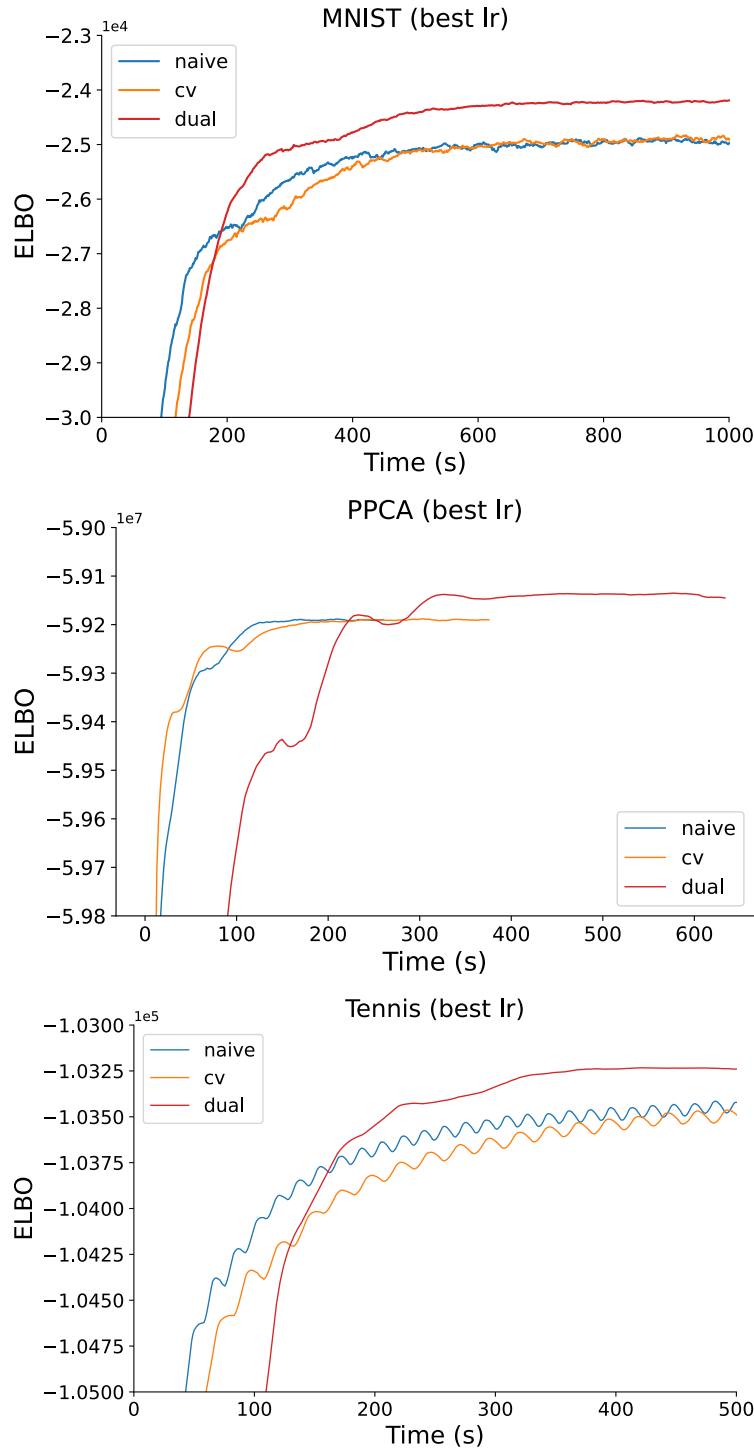


Figure 7: **On large scale problems, the dual estimator allows faster convergence in terms of wall-clock time.** For example, on MNIST, it takes dual around 300 seconds to reach an ELBO of -2.5×10^4 whereas the cv and the naive estimator would take around 600 seconds. On PPCA, dual shows slower convergence in the beginning as SAGA is still warming up, however it is capable of reaching much better results at the end of the optimization.