

A Dual Form of Bregman Momentum

The dual form of Bregman momentum given in (10) can be obtained by first forming the dual Bregman divergence in terms of the dual variables $\mathbf{w}^*(t)$ and \mathbf{w}_0^* and then taking the time derivative:

$$\begin{aligned}\dot{D}_F(\mathbf{w}(t), \mathbf{w}_0) &= \dot{D}_{F^*}(\mathbf{w}_0^*, \mathbf{w}^*(t)) = \frac{\partial}{\partial t} \left(F^*(\mathbf{w}_0^*) - F^*(\mathbf{w}^*(t)) - f^*(\mathbf{w}^*(t))^\top (\mathbf{w}_0^* - \mathbf{w}^*(t)) \right) \\ &= -\dot{F}^*(\mathbf{w}^*(t)) + f^*(\mathbf{w}^*(t))^\top \dot{\mathbf{w}}^*(t) + (\mathbf{w}^*(t) - \mathbf{w}_0^*)^\top \mathbf{H}_{F^*}(\mathbf{w}^*(t)) \dot{\mathbf{w}}^*(t) \\ &= (\mathbf{w}^*(t) - \mathbf{w}_0^*)^\top \mathbf{H}_{F^*}(\mathbf{w}^*(t)) \dot{\mathbf{w}}^*(t),\end{aligned}$$

where we use the fact that $\dot{F}^*(\mathbf{w}^*(t)) = f^*(\mathbf{w}^*(t))^\top \dot{\mathbf{w}}^*(t)$.

B Constrained Updates and Reparameterization

We first provide a proof for Proposition 1. Then, we prove Theorem 3.

Proposition 1. *The CMD update with the additional constraint $\psi(\mathbf{w}(t)) = \mathbf{0}$ for some function $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^m$ s.t. $\{\mathbf{w} \in \mathcal{C} \mid \psi(\mathbf{w}(t)) = \mathbf{0}\}$ is non-empty, amounts to the projected gradient update*

$$\dot{f}(\mathbf{w}(t)) = -\mathbf{P}_\psi(\mathbf{w}(t)) \nabla L(\mathbf{w}(t)) \quad \& \quad \dot{f}^*(\mathbf{w}^*(t)) = -\mathbf{P}_\psi(\mathbf{w}(t))^\top \nabla L \circ f^*(\mathbf{w}^*(t)), \quad (14)$$

where $\mathbf{P}_\psi := \mathbf{I}_d - \mathbf{J}_\psi^\top (\mathbf{J}_\psi \mathbf{H}_F^{-1} \mathbf{J}_\psi^\top)^{-1} \mathbf{J}_\psi \mathbf{H}_F^{-1}$ is the projection matrix onto the tangent space of F at $\mathbf{w}(t)$ and $\mathbf{J}_\psi(\mathbf{w}(t))$. Equivalently, the update can be written as a projected natural gradient descent update

$$\dot{\mathbf{w}}(t) = -\mathbf{P}_\psi^\top(\mathbf{w}(t)) \mathbf{H}_F^{-1}(\mathbf{w}(t)) \nabla L(\mathbf{w}(t)) \quad \& \quad \dot{\mathbf{w}}^*(t) = -\mathbf{P}_\psi \mathbf{H}_{F^*}^{-1}(\mathbf{w}^*(t)) \nabla L \circ f^*(\mathbf{w}^*(t)). \quad (15)$$

Proof of Proposition 1. We use a Lagrange multiplier $\lambda(t) \in \mathbb{R}^m$ in (6) to enforce the constraint $\psi(\mathbf{w}(t)) = \mathbf{0}$ for all $t \geq 0$,

$$\min_{\mathbf{w}(t)} \left\{ \dot{D}_F(\mathbf{w}(t), \mathbf{w}_s) + L(\mathbf{w}(t)) + \lambda(t)^\top \psi(\mathbf{w}(t)) \right\}. \quad (23)$$

Setting the derivative w.r.t. $\mathbf{w}(t)$ to zero, we have

$$\dot{f}(\mathbf{w}(t)) + \nabla_{\mathbf{w}} L(\mathbf{w}(t)) + \mathbf{J}_\psi(\mathbf{w}(t))^\top \lambda(t) = \mathbf{0}, \quad (24)$$

where $\mathbf{J}_\psi(\mathbf{w}(t))$ is the Jacobian of the function $\psi(\mathbf{w}(t))$. In order to solve for $\lambda(t)$, first note that $\dot{\psi}(\mathbf{w}(t)) = \mathbf{J}_\psi(\mathbf{w}(t)) \dot{\mathbf{w}}(t) = \mathbf{0}$. Using the equality $\dot{f}(\mathbf{w}(t)) = \mathbf{H}_F(\mathbf{w}(t)) \dot{\mathbf{w}}(t)$ and multiplying both sides by $\mathbf{J}_\psi(\mathbf{w}(t)) \mathbf{H}_F^{-1}(\mathbf{w}(t))$ yields (ignoring t)

$$\mathbf{J}_\psi(\mathbf{w}) \dot{\mathbf{w}} + \mathbf{J}_\psi(\mathbf{w}) \mathbf{H}_F^{-1}(\mathbf{w}) \nabla L(\mathbf{w}) + \mathbf{J}_\psi(\mathbf{w}) \mathbf{H}_F^{-1}(\mathbf{w}) \mathbf{J}_\psi^\top(\mathbf{w}) \lambda(t) = \mathbf{0}.$$

Assuming that the inverse exists, then

$$\lambda = -(\mathbf{J}_\psi(\mathbf{w}) \mathbf{H}_F^{-1}(\mathbf{w}) \mathbf{J}_\psi^\top(\mathbf{w}))^{-1} \mathbf{J}_\psi(\mathbf{w}) \mathbf{H}_F^{-1}(\mathbf{w}) \nabla L(\mathbf{w}).$$

Plugging in for $\lambda(t)$ yields (15). Multiplying both sides by $\mathbf{H}_F(\mathbf{w})$ and using $\dot{f}(\mathbf{w}) = \mathbf{H}_F(\mathbf{w}) \dot{\mathbf{w}}$ yields (14). \square

Theorem 3. *The constrained CMD update (14) coincides with the reparameterized projected gradient update on the composite loss,*

$$\dot{g}(\mathbf{u}(t)) = -\mathbf{P}_{\psi \circ q}(\mathbf{u}(t)) \nabla_{\mathbf{u}} L \circ q(\mathbf{u}(t)),$$

where $\mathbf{P}_{\psi \circ q} := \mathbf{I}_k - \mathbf{J}_{\psi \circ q}^\top (\mathbf{J}_{\psi \circ q} \mathbf{H}_G^{-1} \mathbf{J}_{\psi \circ q}^\top)^{-1} \mathbf{J}_{\psi \circ q} \mathbf{H}_G^{-1}$ is the projection matrix onto the tangent space at $\mathbf{u}(t)$ and $\mathbf{J}_{\psi \circ q}(\mathbf{u}) := \mathbf{J}_q^\top(\mathbf{u}) \mathbf{J}_\psi(\mathbf{w})$.

Proof of Theorem 3 Similar to the proof of Proposition 1, we use a Lagrange multiplier $\lambda(t) \in \mathbb{R}^m$ to enforce the constraint $\psi \circ q(\mathbf{u}(t)) = \mathbf{0}$ for all $t \geq 0$,

$$\min_{\mathbf{u}(t)} \left\{ D_G(\mathbf{u}(t), \mathbf{u}_s) + L \circ q(\mathbf{u}(t)) + \lambda(t)^\top \psi \circ q(\mathbf{u}(t)) \right\}.$$

Setting the derivative w.r.t. $\mathbf{u}(t)$ to zero, we have

$$\dot{g}(\mathbf{w}(t)) + \nabla_{\mathbf{u}} L \circ q(\mathbf{w}(t)) + \mathbf{J}_{\psi \circ q}^\top(\mathbf{u}(t)) \lambda(t) = \mathbf{0},$$

where $\mathbf{J}_{\psi \circ q}(\mathbf{u}(t)) := \mathbf{J}_q^\top(\mathbf{u}) \nabla \psi(\mathbf{w}(t))$. In order to solve for $\lambda(t)$, we use the fact that $\dot{\psi} \circ q(\mathbf{u}(t)) = \mathbf{J}_{\psi \circ q}(\mathbf{u}(t)) \dot{\mathbf{u}}(t) = \mathbf{0}$. Using the equality $\dot{g}(\mathbf{u}(t)) = \mathbf{H}_G(\mathbf{u}(t)) \dot{\mathbf{u}}(t)$ and multiplying both sides by $\mathbf{J}_{\psi \circ q}(\mathbf{u}(t)) \mathbf{H}_G^{-1}(\mathbf{u}(t))$ yields (ignoring t)

$$\mathbf{J}_{\psi \circ q}(\mathbf{u}) \dot{\mathbf{u}} + \mathbf{J}_{\psi \circ q}(\mathbf{w}) \mathbf{H}_G^{-1}(\mathbf{u}) \nabla L \circ q(\mathbf{u}) + \mathbf{J}_{\psi \circ q}(\mathbf{w}) \mathbf{H}_G^{-1}(\mathbf{w}) \mathbf{J}_{\psi \circ q}^\top(\mathbf{u}) \lambda(t) = \mathbf{0}.$$

The rest of the proof follows similarly by solving for $\lambda(t)$ and rearranging the terms. Finally, applying the results of Theorem 2 concludes the proof. \square

C Discretized Updates

In this section, we discuss different strategies for discretizing the CMD updates and provide examples for each case.

The most straight-forward discretization of the unconstrained CMD update (1) is the forward Euler (i.e. explicit) discretization, given in (5). Note that this corresponds to a (approximate) minimizer of the discretized form of (6) (with initial condition at \mathbf{w}_s instead of \mathbf{w}_0), that is,

$$\operatorname{argmin}_{\mathbf{w}} \left\{ 1/h \left(D_F(\mathbf{w}, \mathbf{w}_s) - \underbrace{D_F(\mathbf{w}_s, \mathbf{w}_s)}_{=0} \right) + L(\mathbf{w}) \right\}.$$

An alternative way of discretizing is to apply the approximation on the equivalent natural gradient form (11), which yields

$$\mathbf{w}_{s+1} - \mathbf{w}_s = -h \mathbf{H}_F^{-1}(\mathbf{w}_s) \nabla L(\mathbf{w}_s).$$

Despite being equivalent in continuous-time, the two approximations may correspond to different updates after discretization. As an example, for the EGU update motivated by $f(\mathbf{w}) = \log \mathbf{w}$ link, the latter approximation yields

$$\mathbf{w}_{s+1} = \mathbf{w}_s \odot (\mathbf{1} - h \nabla L(\mathbf{w}_s)),$$

which amounts to approximating the exponential factor $\exp(-\eta \nabla L(\mathbf{w}_s))$ in the EGU update by its Taylor expansion $(\mathbf{1} - h \nabla L(\mathbf{w}_s))$.

The situation becomes more involved for discretizing the constrained updates. As the first approach, it is possible to directly discretize the projected CMD update (14)

$$f(\tilde{\mathbf{w}}_{s+1}) - f(\mathbf{w}_s) = -h \mathbf{P}_\psi(\mathbf{w}_s) \nabla L(\mathbf{w}_s).$$

However, note that the new parameter $\tilde{\mathbf{w}}_{s+1}$ may fall outside the constraint set $\mathcal{C}_\psi := \{\mathbf{w} \in \mathcal{C} \mid \psi(\mathbf{w}) = \mathbf{0}\}$. As a result, a Bregman projection [Shalev-Shwartz et al., 2012] into \mathcal{C}_ψ may need to be applied after the update, that is

$$\mathbf{w}_{s+1} = \operatorname{argmin}_{\mathbf{w} \in \mathcal{C}_\psi} D_F(\mathbf{w}, \tilde{\mathbf{w}}_{s+1}). \quad (25)$$

As an example, for the normalized EG updates with the additional constraint that $\mathbf{w}^\top \mathbf{1} = 1$, we have $\mathbf{P}_\psi(\mathbf{w}) = \mathbf{I}_d - \mathbf{1} \mathbf{w}^\top$ and the approximation yields

$$\log(\tilde{\mathbf{w}}_{s+1}) - \log(\mathbf{w}_s) = -h (\nabla L(\mathbf{w}_s) - \mathbf{1} \mathbb{E}_{\mathbf{w}_s}[\nabla L(\mathbf{w}_s)]),$$

where $\mathbb{E}_{\mathbf{w}_s}[\nabla L(\mathbf{w}_s)] = \mathbf{w}_s^\top \nabla L(\mathbf{w}_s)$. Clearly, $\tilde{\mathbf{w}}_{s+1}$ may not necessarily satisfy $\tilde{\mathbf{w}}_{s+1}^\top \mathbf{1} = 1$. Therefore, we apply

$$\mathbf{w}_{s+1} = \frac{\tilde{\mathbf{w}}_{s+1}}{\|\tilde{\mathbf{w}}_{s+1}\|_1},$$

which corresponds to the Bregman projection onto the unit simplex using the relative entropy divergence [Kivinen and Warmuth, 1997].

An alternative approach for discretizing the constrained update would be to first discretize the functional objective with the Lagrange multiplier (23) and then (approximately) solve for the update. That is,

$$\mathbf{w}_{s+1} = \operatorname{argmin}_{\mathbf{w}} \left\{ 1/h \left(D_F(\mathbf{w}, \mathbf{w}_s) - \underbrace{D_F(\mathbf{w}_s, \mathbf{w}_s)}_{=0} \right) + L(\mathbf{w}) + \boldsymbol{\lambda}^\top \psi(\mathbf{w}) \right\}.$$

Note that in this case, the update satisfies the constraint $\psi(\mathbf{w}_{s+1}) = \mathbf{0}$ because of directly using the Lagrange multiplier. For the normalized EG update, this corresponds to the original normalized EG update in [Littlestone and Warmuth, 1994],

$$\mathbf{w}_{s+1} = \frac{\mathbf{w}_s \odot \exp(-h \nabla L(\mathbf{w}_s))}{\|\mathbf{w}_s \odot \exp(-h \nabla L(\mathbf{w}_s))\|_1}.$$

Finally, it is also possible to discretize the projected natural gradient update (15). Again, a Bregman projection into \mathcal{C}_ψ may need to be required after the update, that is,

$$\tilde{\mathbf{w}}_{s+1} - \mathbf{w}_s = -h \mathbf{P}_\psi(\mathbf{w}_s)^\top \mathbf{H}_F^{-1}(\mathbf{w}_s) \nabla L(\mathbf{w}(t)),$$

followed by (25). For the normalized EG update, the first step corresponds to

$$\mathbf{w}_{s+1} = \mathbf{w}_s \odot \left(\mathbf{1} - h(\nabla L(\mathbf{w}_s) - \mathbf{1} \mathbb{E}_{\mathbf{w}_s}[\nabla L(\mathbf{w}_s)]) \right),$$

which recovers to the *approximated EG* update of [Kivinen and Warmuth 1997]. Note that $\mathbf{w}_{s+1}^\top \mathbf{1} = 1$ and therefore, no projection step is required in this case.