A EXAMPLES OF $\mathcal{K}$

We provide a list of examples of $\mathcal{K}$ and the corresponding $\nabla\mathcal{K}$ and $\mathcal{K}^*$. It is useful to define the following indicator functions of set $\{z = 0\}$:

$$\delta(z) = \begin{cases} 0 & \text{if } z = 0 \\ +\infty & \text{if } z \neq 0. \end{cases}, \qquad \mathbb{I}(z) = \begin{cases} 0 & \text{if } z = 0 \\ 1 & \text{if } z \neq 0. \end{cases},$$

Note that δ is the conjugate function of $f(x) = x$, as $\delta(x) = \sup_z x^\top z$.

ℓ_p norm When $\mathcal{K}(x) = \|x\|_p = (\sum_i |x_i|^p)^{1/p}$ for $p \geq 1$, we can take

$$\nabla\mathcal{K}(x) = \frac{\operatorname{sign}(x) |x|^{p-1}}{\|x\|_p^{p-1}},$$

and

$$\mathcal{K}^*(x) = \sup_z x^\top z - \|z\|_p = \sup_{c \geq 0} \|x\|_q c - c = \delta(\|x\|_q \leq 1),$$

where q is the conjugate number of p, satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Hence, Lion-$\mathcal{K}$ with ℓ_p norm correspond to solving

$$\min_x f(x) \quad s.t. \quad \|x\|_q \leq 1/\lambda.$$

Group ℓ_p norm Assume x is partitioned into a number of groups: $x = [x_{\mathcal{G}_i}]_{i=1}^k$. Consider the group ℓ_p norm: $\mathcal{K}(x) = \sum_{i=1}^k \|x_{\mathcal{G}_i}\|_p$. Then, we can take

$$\nabla\mathcal{K}(x) = \left[\frac{\operatorname{sign}(x_{\mathcal{G}_i}) |x_{\mathcal{G}_i}|^{p-1}}{\|x_{\mathcal{G}_i}\|_p^{p-1}}\right]_{i=1}^k$$

The conjugate function is

$$\mathcal{K}^*(x) = \sup_z \sum_{i=1}^k x_{\mathcal{G}_i}^\top z_{\mathcal{G}_i} - \|z_{\mathcal{G}_i}\|_p = \sum_{i=1}^k \delta(\|x_{\mathcal{G}_i}\|_q \leq 1).$$

Hence, Lion-$\mathcal{K}$ with grouped ℓ_p norm corresponds to solving

$$\min_x f(x) \quad s.t. \quad \|x_{\mathcal{G}_i}\|_q \leq 1/\lambda, \quad \forall i.$$

Lower Truncated ℓ_1 Norm Consider $\mathcal{K}(x) = \sum_{i=1}^d \max(|x_i| - e, 0)$ where $e > 0$. We can take

$$\nabla\mathcal{K}(x) = \mathbb{I}(|x| \geq e)\operatorname{sign}(x), \tag{16}$$

which uses $\operatorname{sign}(x)$ as Lion, but zeros out the gradient on the elements with absolute values smaller than e. The conjugate is

$$\begin{aligned}
\mathcal{K}^*(x) &= \sup_z \sum_{i=1}^d (x_i z_i - \max(|z|_i - e, 0)) \\
&= \sup_{z,c} \sum_{i=1}^d (x_i z_i - c_i) \quad s.t. \quad c_i \geq 0, \quad c \geq |z_i| - e \\
&= \sup_{c \geq 0} \sum_{i=1}^d |x_i| (c_i + e) - c_i \\
&= \sum_{i=1}^d \delta(|x_i| \leq 1) + e |x_i| \\
&= \delta(\|x\|_\infty \leq 1) + e \|x\|_1 .
\end{aligned}$$

Hence, Lion-$\mathcal{K}$ corresponds to solving

$$\min_x \alpha f(x) + e\gamma \|x\|_1 \quad s.t. \quad \|x\|_\infty \leq 1/\lambda. \tag{17}$$

Hence, truncating the small gradients in Lion induces an ℓ_1 penalty, which encourages the sparsity of the final solution.

Lower (Vector-wise) Truncated ℓ_p Norm Consider $\mathcal{K}(x) = \max(\|x\|_p - e, 0)$. We have

$$\nabla \mathcal{K}(x) = \mathbb{I}(\|x\|_p - e \geq 0) \frac{\mathrm{sign}(x)\,|x|^{p-1}}{\|x\|_p^{p-1}},$$

in which the gradient is zeroed out when $\|x\|_p \leq e$. The conjugate is

$$\begin{aligned}
\mathcal{K}^*(x) &= \sup_z (x^\top z - \max(\|z\|_p - e, 0)) \\
&= \sup_{z,c} (x^\top z - c) \quad s.t. \quad c \geq 0, \quad c \geq \|z\|_p - e \\
&= \sup_{c \geq 0} \|x\|_q (c + e) - c \\
&= \delta(\|x\|_q \leq 1) + e \|x\|_q .
\end{aligned}$$

Hence, Lion-$\mathcal{K}$ corresponds to solving

$$\min_x \alpha f(x) + e\gamma \|x\|_q \qquad s.t. \qquad \|x\|_q \leq 1/\lambda.$$

Sorting Norm For $x = [x_1, \ldots, x_d]$, let $|x_{(1)}| \geq |x_{(2)}| \ldots$ be the sorting of the elements by absolute values. Define

$$\text{Sorting norm:} \quad \mathcal{K}(x) = \sum_i c_i \left|x_{(i)}\right|,$$

where $c_1 \geq c_2 \geq \ldots \geq 0$ is a descending non-negative sequence. The sorting norm is convex because it can be represented as the supreme of a set of convex functions, by the rearrangement inequality, as follows

$$\mathcal{K}(x) = \max_{\sigma \in \Gamma} \sum_{i=1}^d c_{\sigma(i)} |x_i|,$$

where Γ denotes the set of permutations on $\{1, \ldots, n\}$. One subgradient of $\mathcal{K}$ is

$$\nabla \mathcal{K}(x)_i = c_{rank(i,x)} \mathrm{sign}(x_i),$$

where $rank(i, x)$ denotes the rank of $|x_i|$ in x.

$$\begin{aligned}
\mathcal{K}^*(x) &= \sup_z \left\{ x^\top z - \sum_i c_i \left|z_{(i)}\right| \right\} \\
&= \sup_{z \geq 0} \left\{ \sum_i \left|x_{(i)}\right| \times z_{(i)} - \sum_i c_i z_{(i)} \right\} && \text{//by rearrangement inequality} \\
&= \sup_{w \geq 0} \left\{ \sum_i (\left|x_{(i)}\right| - c_i) \times (\sum_{j \geq i} w_j) \right\} && \text{//let } z_{(i)} = \sum_{j \geq i} w_j, w_j \geq 0 \\
&= \sup_{w \geq 0} \left\{ \sum_j \sum_{i \leq j} (\left|x_{(i)}\right| - c_i) \times w_j \right\} && \text{//let } z_{(i)} = \sum_{j \geq i} w_j, w_j \geq 0 \\
&= \sum_j \delta(\sum_{i \leq j} \left|x_{(i)}\right| \leq \sum_{j \leq i} c_j)
\end{aligned}$$

Hence, Lion-$\mathcal{K}$ corresponds to imposing a sequence of bounds on the cumsum of the sorted x:

$$\min_x f(x) \quad s.t. \quad \sum_{j \leq i} \left|x_{(i)}\right| \leq C_i, \quad \text{where } C_i = \sum_{j \leq i} c_j.$$

An interesting special case is when $c_i = \mathbb{I}(i \leq i^{cut})$ for some integer $i^{cut} \in \{1, \ldots, d\}$, so that

$$\mathcal{K}(x) = \sum_{i \leq i^{cut}} \left|x_{(i)}\right|, \qquad \nabla \mathcal{K}(x) = \mathbb{I}(|x| \geq x_{(i^{cut})}) \mathrm{sign}(x),$$

in which we zero out the updates of the elements whose absolute values are smaller than the i^{cut}-th largest element. It is useful to compare this with (16) which applies the truncation based on a fixed number ϵ, rather than the percentile.

The conjugate is

$$\mathcal{K}^*(x) = \sum_{j \le i^{cut}} \delta(\left|x_{(j)}\right| \le 1) + \delta(\|x\|_1 \le i^{cut})$$

Then, Lion-$\mathcal{K}$ in this case corresponds to solving

$$\min_x f(x) \quad s.t. \quad \|x\|_1 \le i^{cut}/\lambda, \qquad \|x\|_\infty \le 1/\lambda,$$

in which the percentile-based truncation effectively imposes a constraint on the ℓ_1 norm of x. It is different from (17) in which the ℓ_1 norm appears as a regularization term in the objective, rather than as a hard constraint.

Entropy Consider $\mathcal{K}(x) = \sum_{i=1}^d \frac{1}{a} \log\left(\frac{1}{2}(\exp(ax_i) + \exp(-ax_i))\right)$, where $a > 0$. We have

$$\nabla \mathcal{K}(x) = \frac{\exp(ax) - \exp(-ax)}{\exp(ax) + \exp(-ax)} = \tanh(ax).$$

Taking the inverse, we have $\nabla \mathcal{K}^*(x) = \frac{1}{2a} \log \frac{1+x}{1-x}$, with domain in $\|x\|_\infty \le 1$. by integration, the conjugate function is hence,

$$\mathcal{K}^*(x) = \sum_{i=1}^d \frac{1}{2a}(x_i + 1)\log(x_i + 1) + \frac{1}{2a}(1 - x_i)\log(1 - x_i) + \delta(\|x\|_\infty < 1).$$

Lion-$\mathcal{K}$ correspond to solving an entropy-regularized optimization:

$$\min_x \alpha f(x) + \frac{\gamma}{\lambda} E(\lambda x) \quad s.t. \quad \|x\|_\infty \le 1/\lambda,$$

where $E(x) = \sum_{i=1}^d \frac{1}{2a}(x_i + 1)\log(x_i + 1)n + \frac{1}{2a}(1 - x_i)\log(1 - x_i)$.

Huber Loss For $a \ge 0$, define the Huber loss:

$$\mathcal{K}(x) = \sum_{i=1}^d \text{Huber}_a(x_i) \quad \text{where} \quad \text{Huber}_a(x_i) = \mathbb{I}(|x_i| \ge a) \times |x_i| + \mathbb{I}(|x_i| < a) \times \frac{1}{2a} x_i^2,$$

We have

$$\nabla \mathcal{K}(x) = \text{Clip}(x, -a, a)/a, \qquad \text{with} \qquad \text{Clip}(x_i, a, b) = \begin{cases} x_i & \text{if } x \in [a, b] \\ b & \text{if } x > b \\ a & \text{if } x < a. \end{cases}$$

The conjugate is

$$\mathcal{K}^*(x) = \frac{a}{2}\|x\|_2^2 + \delta(\|x\|_\infty \le 1),$$

$$\begin{aligned} \mathcal{K}^*(x) &= \sum_{i=1}^d \max(\sup_{|z| \ge a} x_i z_i - |z_i|, \quad \sup_{|z_i| < a} x_i z_i - \frac{1}{2a} z_i^2) \\ &= \sum_{i=1}^d \max\left(\delta(|x_i| \le 1) + a(|x_i| - 1), \ \frac{1}{2} a x_i^2\right) \\ &= \sum_{i=1}^d \delta(|x_i| \le 1) + \frac{1}{2} a x_i^2 \\ &= \frac{a}{2}\|x\|_2^2 + \delta(\|x\|_\infty \le 1). \end{aligned}$$

Relativistic Consider $\mathcal{K}(x) = \sum_{i=1}^{d} \sqrt{x_i^2 + e^2}$, then $\nabla \mathcal{K}(x) = \frac{x}{\sqrt{x^2+e^2}}$, and

$$\begin{aligned}\mathcal{K}^*(x) &= \sup_z \left(\sum_{i=1}^{d} x_i z_i - \sqrt{z_i^2 + e^2} \right) \\ &= \sum_{i=1}^{d} \frac{x_i^2 e}{\sqrt{1 - x_i^2}} - \frac{e}{\sqrt{1 - x_i^2}} \qquad \text{//Solution: } z_i^2 = \frac{x_i^2 e^2}{1 - x_i^2} \\ &= \sum_{i=1}^{d} -e\sqrt{1 - x_i^2} + \delta(|x_i| \le 1) \\ &= \sum_{i=1}^{d} -e\sqrt{1 - x_i^2} + \delta(\|x\|_\infty \le 1).\end{aligned}$$

A related case is

$$\mathcal{K}(x) = |x| - e\log(|x|/e + 1), \qquad \text{with} \qquad \nabla \mathcal{K}(x) = \frac{x}{|x| + e},$$

whose conjugate function is

$$\begin{aligned}\mathcal{K}^*(x) &= \sup_x \left(\sum_{i=1}^{d} x_i z_i - |z_i| + e\log(|z_i|/e + 1) \right) \\ &= \sum_{i=1}^{d} |x_i|^2 e/(1 - |x_i|) - |x_i| e/(1 - |x_i|) + e\log(1/(1 - |x_i|)) \qquad \text{//Solution: } z = |x| e/(1 - |x|) \\ &= \sum_{i=1}^{d} -e(|x_i| + \log(1 - |x_i|)) + \delta(\|x\|_\infty < 1).\end{aligned}$$

B Proofs

B.1 Convex Function Preliminaries

Lemma 2.1 *Assume $\mathcal{K}, \mathcal{K}^*$ is a closed convex conjugate pair and $\nabla \mathcal{K}$, $\nabla \mathcal{K}$ are their subgradients, we have*

$$(\nabla \mathcal{K}(x) - \nabla \mathcal{K}(y))^\top (x - y) \ge 0, \qquad (\nabla \mathcal{K}(x) - y)^\top (x - \nabla \mathcal{K}^*(y)) \ge 0. \tag{18}$$

Proof. 1) By definition of subgradient, we have

$$\begin{aligned}\mathcal{K}(y) - \mathcal{K}(x) &\ge \nabla \mathcal{K}(x)^\top (y - x) \\ \mathcal{K}(x) - \mathcal{K}(y) &\ge \nabla \mathcal{K}(y)^\top (x - y).\end{aligned}$$

Summing them together yields $(\nabla \mathcal{K}(x) - \nabla \mathcal{K}(y))^\top (x - y) \ge 0$.

2) Because $\nabla \mathcal{K}^*(y) \in \partial \mathcal{K}^*(y)$, we have

$$\mathcal{K}^*(\nabla \mathcal{K}(x)) - \mathcal{K}^*(y) \ge \nabla \mathcal{K}^*(y)^\top (\nabla \mathcal{K}(x) - y),$$

Because $\nabla \mathcal{K}(x) \in \partial \mathcal{K}(x)$, by the property of conjugate functions, we have $x \in \partial \mathcal{K}^*(\nabla \mathcal{K}(x))$, and hence

$$\mathcal{K}^*(y) - \mathcal{K}^*(\nabla \mathcal{K}(x)) \ge x^\top (y - \nabla \mathcal{K}(x)).$$

Summing the two inequalities above yields

$$(\nabla \mathcal{K}(x) - y)^\top (\nabla \mathcal{K}^*(y) - x) \le (\mathcal{K}^*(\nabla \mathcal{K}(x)) - \mathcal{K}^*(y)) + (\mathcal{K}^*(y) - \mathcal{K}^*(\nabla \mathcal{K}(x))) = 0.$$

□

B.2 Connection with Nesterov Momentum

Lemma B.1. *The Lion-$\mathcal{K}$ ODE is*

$$\begin{aligned}\dot{x}_t &= \nabla\mathcal{K}(m_t - \varepsilon(\alpha\nabla f(x_t) + \gamma m_t)) - \lambda x_t \\ \dot{m}_t &= -\alpha\nabla f(x_t) - \gamma m_t.\end{aligned}$$

is equivalent to

$$\nabla^2\mathcal{K}^*(\dot{x}_t + \lambda x_t)(\ddot{x}_t + \lambda\dot{x}_t) + \varepsilon\alpha\nabla^2 f(x_t)\dot{x}_t + \gamma\nabla\mathcal{K}^*(\dot{x}_t + \lambda x_t) + \alpha\nabla f(x_t) = 0, \tag{19}$$

if $\mathcal{K}^*$ and f are second order differentiable.

In particular, if $\mathcal{K}(x) = \|x\|_2^2/2$, we have

$$\ddot{x}_t + (\lambda+\gamma)\dot{x}_t + \varepsilon\alpha\nabla^2 f(x_t)\dot{x}_t + \gamma\lambda x_t + \alpha\nabla f(x_t) = 0. \tag{20}$$

This ODE minimizes $F(x) = \alpha f(x) + \gamma\lambda\|x\|_2^2/2$.

Remark We have the following observations from (21):

1) The role of the weight decay λ and momentum damping coefficient γ is symmetric in (21).

2) When either the weight decay or momentum damping is turned off, i.e., $\gamma\lambda = 0$, the ℓ_2 regularization in $F(x)$ is turned off, and we have

$$\ddot{x}_t + (\lambda+\gamma)\dot{x}_t + \varepsilon\alpha\nabla^2 f(x_t)\dot{x}_t + \alpha\nabla f(x_t) = 0, \tag{21}$$

which coincides with the *high-resolution* ODE [35] that serves as a continuous-time modeling of Nesterov momentum for minimizing $f(x)$.

3) The Hessian-dependent damping term $\nabla^2 f(x_t)\dot{x}_t$ arises to due the gradient enhancement ($\varepsilon > 0$), and it is known to play a key role in Nesterov momentum and acceleration [1, 35]. When we turn off the gradient enhancement ($\varepsilon = 0$), we get

$$\ddot{x}_t + (\lambda+\gamma)\dot{x}_t + \alpha\nabla f(x_t) = 0,$$

which is the ODE for Polayk momentum, the equation of motion of a ball with unit mass moving in a potential field $\alpha f(x)$ with a friction coefficient $(\lambda+\gamma)$.

Proof. We want to cancel out m_t. The first equation yields

$$(1-\varepsilon\gamma)m_t = \left(\nabla\mathcal{K}^*(\dot{x}_t + \lambda x_t) + \varepsilon\alpha\nabla f(x_t)\right). \tag{22}$$

Plugging it into the second equation yields

$$\begin{aligned}(1-\varepsilon\gamma)\dot{m}_t &= -\alpha(1-\varepsilon\gamma)\nabla f(x_t) - \gamma\left(\nabla\mathcal{K}^*(\dot{x}_t + \lambda x_t) + \varepsilon\alpha\nabla f(x_t)\right) \\ &= -\alpha\nabla f(x_t) - \gamma\nabla\mathcal{K}^*(\dot{x}_t + \lambda x_t).\end{aligned} \tag{23}$$

Combining (22) and (23) yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\nabla\mathcal{K}^*(\dot{x}_t + \lambda x_t) + \varepsilon\alpha\nabla f(x_t)\right) = -\alpha\nabla f(x_t) - \gamma\nabla\mathcal{K}^*(\dot{x}_t + \lambda x_t).$$

Or

$$\nabla^2\mathcal{K}^*(\dot{x}_t + \lambda x_t)(\ddot{x}_t + \lambda\dot{x}_t) + \varepsilon\alpha\nabla^2 f(x_t)\dot{x}_t + \gamma\nabla\mathcal{K}^*(\dot{x}_t + \lambda x_t) + \alpha\nabla f(x_t) = 0.$$

□

B.3 Discrete-time Schemes of Lion-$\mathcal{K}$

In the most general form, the Euler approximation of the Lion-$\mathcal{K}$ ODE with step size ϵ is

$$\begin{aligned}x_{t+1} &= x_t + \epsilon(\nabla\mathcal{K}(m_t - \varepsilon(\alpha\nabla f(x_t) + \gamma m_t)) - \lambda x_t) \\ m_{t+1} &= m_t - \epsilon(\alpha\nabla f(x_t) + \gamma m_t),\end{aligned} \tag{24}$$

The discrete Lion-$\mathcal{K}$ scheme in (2) is recovered when $\alpha = \gamma$, $\beta_1 = 1 - \varepsilon\gamma$, $\beta_2 = 1 - \epsilon\gamma$. By scaling $f(x)$ by a positive multiplicative ratio, (2) in fact covers all cases of (24) when $\gamma \neq 0$.

When $\gamma = 0$, however, (24) reduces to a momentum-undamped variant of Lion-$\mathcal{K}$:

$$\text{Undamped Lion-}\mathcal{K}\text{:} \qquad \begin{aligned} x_{t+1} &= x_t + \epsilon(\nabla\mathcal{K}(m_t - \beta_1 \nabla f(x_t)) - \lambda x_t) \\ m_{t+1} &= m_t - \beta_2 \nabla f(x_t), \end{aligned}$$

which is the Euler approximation of Lion-$\mathcal{K}$ ODE $\gamma = 0$, step size ϵ, and $\beta_1 = \varepsilon\alpha$, and $\beta_2 = \epsilon\alpha$. Due to $\gamma = 0$, the undamped Lion-$\mathcal{K}$ amounts to solving $\min_x f(x)$, without the regularization $\mathcal{K}^*(\lambda x)$.

The connection to Polyak and Nesterov momentum discussed in Section extends to discrete-time forms. From the first equation (24), we have

$$m_t = \frac{1}{1 - \varepsilon\gamma}\left(\nabla\mathcal{K}^*\left(\frac{x_{t+1} - x_t}{\epsilon} + \lambda x_t\right) + \varepsilon\alpha\nabla f(x_t)\right).$$

Plugging it into the second equation of (24), we get

$$\left(\nabla\mathcal{K}^*\left(\frac{x_{t+2} - x_{t+1}}{\epsilon} + \lambda x_{t+1}\right) + \varepsilon\alpha\nabla f(x_{t+1})\right) = (1 - \epsilon\gamma)\left(\nabla\mathcal{K}^*\left(\frac{x_{t+1} - x_t}{\epsilon} + \lambda x_t\right) + \varepsilon\alpha\nabla f(x_t)\right) - (1 - \varepsilon\gamma)\epsilon\alpha\nabla f(x_t).$$

Hence,

$$\nabla\mathcal{K}^*\left(\frac{x_{t+2} - x_{t+1}}{\epsilon} + \lambda x_{t+1}\right) = -\varepsilon\alpha\nabla f(x_{t+1}) + (1-\epsilon\gamma)\nabla\mathcal{K}^*\left(\frac{x_{t+1} - x_t}{\epsilon} + \lambda x_t\right) + (\varepsilon - \epsilon)\alpha\nabla f(x_t).$$

When $\nabla\mathcal{K}^*(x) = x$, we have

$$x_{t+2} = (1 - \epsilon\lambda)x_{t+1} - \epsilon\varepsilon\alpha\nabla f(x_{t+1}) + (1 - \epsilon\gamma)((x_{t+1} - x_t) + \epsilon\lambda x_t) + \epsilon(\varepsilon - \epsilon)\alpha\nabla f(x_t).$$

It is simplified into

$$x_{t+2} = (1 - \epsilon^2\lambda\gamma)x_{t+1} - \epsilon^2\alpha\nabla f(x_{t+1}) + (1 - \epsilon\gamma)(1 - \epsilon\lambda)(x_{t+1} - x_t) - \epsilon(\varepsilon - \epsilon)\alpha(\nabla f(x_{t+1}) - \nabla f(x_t)).$$

When $\varepsilon > \epsilon$ (corresponding to $\beta_1 < \beta_2$ in Lion-$\mathcal{K}$ (2)), this can be shown to be identical to the Nesterov momentum algorithm for minimizing $F(x) = \alpha f(x) + \lambda\gamma \|x\|_2^2/2$. When $\varepsilon = \epsilon$ (corresponding to $\beta_1 = \beta_2$ in (2)), it is identical to Polyak momentum.

B.4 Frank-Wolfe and Mirror Descent

Frank-Wolfe When $\varepsilon\gamma = 1$, Lion-$\mathcal{K}$ reduces to

$$\dot{x}_t = \nabla\mathcal{K}(-\nabla f(x_t)) - \lambda x_t, \tag{25}$$

where we also set $\varepsilon\alpha = 1$ without loss of generality. In this case, the ODE monotonically decreases the objective

$$F(x) = f(x) + \frac{1}{\lambda}\mathcal{K}^*(\lambda x),$$

without resorting to an additional Lyapunov function. This can be seen from

$$\frac{\mathrm{d}}{\mathrm{d}t}F(x_t) = (\nabla f(x) + \nabla\mathcal{K}^*(\lambda x))^\top(\nabla\mathcal{K}(-\nabla f(x)) - \lambda x) \leq 0,$$

where the inequality follows Lemma 2.1.

The Euler discretization of (25) is

$$x_{t+1} = x_t + \epsilon\left(\nabla\mathcal{K}(-\nabla f(x_t)) - \lambda x_t\right). \tag{26}$$

This can also be derived from conditional gradient descent, or Frank–Wolfe. To see this, recall that the conditional gradient descent update for the $F(x)$ above is

$$\begin{aligned} y_{t+1} &= \arg\min_x \left\{\nabla f(x_t)^\top(x - x_t) + \frac{1}{\lambda}\mathcal{K}^*(\lambda x)\right\} \\ x_{t+1} &= x_t + \epsilon_0(y_{t+1} - x_t), \end{aligned}$$

Solving y_{t+1} yields

$$y_{t+1} = \frac{1}{\lambda}\nabla\mathcal{K}(-\nabla f(x_t)), \qquad \text{and hence} \qquad x_{t+1} = (1 - \epsilon_0)x_t + \frac{\epsilon_0}{\lambda}\nabla\mathcal{K}(-\nabla f(x_t)).$$

Taking $\epsilon = \epsilon_0\lambda$ yields (26).

Dual Space Preconditioning and Mirror Descent When we further set $\lambda = 0$ in (26), Lion-$\mathcal{K}$ reduces to

$$x_{t+1} = x_t + \epsilon \nabla \mathcal{K}(-\nabla f(x_t)), \tag{27}$$

When $\nabla \mathcal{K}(0) = 0$, Eq. (27) is dual space preconditioning [23], which is closely related to mirror descent [26], for minimizing $f(x)$. To see the connection with mirror descent, note that (27) is equivalent to

$$x_{t+1} = x_t + \epsilon \delta_t, \qquad \text{with} \qquad \delta_t = \arg\min_{\delta} \left\{ \nabla f(x_t)^\top \delta + \mathcal{K}^*(\delta) \right\}.$$

Because $\mathcal{K}^*$ and $\mathcal{K}$ are differentiable, then $\nabla \mathcal{K}(0) = 0$ implies $\nabla \mathcal{K}^*(0) = 0$, and hence $\mathcal{K}^*$ achieves the minimum at zero. In this case, $\mathcal{K}^*(\delta) - \mathcal{K}^*(0)$ can be viewed as a Bregman divergence, and hence justifying the connection of (27) with mirror descent. Recall that the Bregman divergence $B_h(x \,||\, y)$ is the Bregman divergence associated with a convex function $h\colon \mathbb{R}^d \to \mathbb{R}$ is defined as

$$B_h(x \,||\, y) = h(x) - h(y) - \nabla h(y)^\top (x - y).$$

With $\nabla \mathcal{K}^*(0) = 0$, it is then easy to show

$$\mathcal{K}^*(\delta) - \mathcal{K}^*(0) = B_{\mathcal{K}^*}(\delta \,||\, 0) = B_{\mathcal{K}_t^*}(x_t + \epsilon\delta \,||\, x_t),$$

where $\mathcal{K}_t^* = \mathcal{K}^*\left(\frac{x - x_t}{\epsilon}\right)$.

B.5 Lion-$\mathcal{K}$ without Gradient Enhancement ($\varepsilon = 0$)

Theorem B.2. *Consider the ODE of Lion-$\mathcal{K}$-W without gradient correction:*

$$\begin{aligned} \dot{x}_t &= \nabla \mathcal{K}(m_t) - \lambda x_t \\ \dot{m}_t &= -\alpha \nabla f(x_t) - \gamma m_t, \end{aligned} \tag{28}$$

with $\lambda, \alpha, \gamma > 0$. Its fixed point is the minimum of

$$\min_x \alpha f(x) + \frac{\gamma}{\lambda} \mathcal{K}^*(\lambda x).$$

It yields the following Lyapunov function:

$$H(x, m) = \alpha f(x) + \frac{\gamma}{\lambda} \mathcal{K}^*(\lambda x) + (\mathcal{K}^*(\lambda x) + \mathcal{K}(m) - \lambda x^\top m).$$

Proof. Observe that

$$\begin{aligned} \nabla_x H(x, m) &= \alpha \nabla f(x) + (\gamma + \lambda) \nabla \mathcal{K}^*(\lambda x) - \lambda m \\ \nabla_m H(x, m) &= \nabla \mathcal{K}(m) - \lambda x, \end{aligned}$$

and (28) can be written into

$$\begin{aligned} \dot{x}_t &\coloneqq V_x(x_t, m_t) = \nabla_m H(x_t, m_t) \\ \dot{m}_t &\coloneqq V_m(x_t, m_t) = -\nabla_x H(x_t, m_t) - \hat{H}_m(x_t, m_t), \end{aligned}$$

with $\hat{H}_m(x_t, m_t) = (\gamma + \lambda)(m_t - \nabla \mathcal{K}^*(\lambda x_t))$. By Lemma 2.1, we have

$$\hat{H}_m^\top (\nabla_m H) = (m - \nabla \mathcal{K}^*(\lambda x))^\top (\nabla \mathcal{K}(m) - \lambda x) \geq 0.$$

Then

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} H(x_t, m_t) &= \nabla_x H^\top V_x + \nabla_m H^\top V_m \\ &= \nabla_x H^\top (\nabla_m H) + \nabla_m H^\top (-\nabla_x H - \hat{H}_m) = -\nabla_m H^\top \hat{H}_m \leq 0. \end{aligned}$$

In fact, this ODE has a Hamiltonian + descent structure [22], as it can viewed as a Hamiltonian system damped with a descending force:

$$\begin{bmatrix} \dot{x}_t \\ \dot{m}_t \end{bmatrix} = \underbrace{\begin{bmatrix} +\nabla_m H(x_t, m_t) \\ -\nabla_x H(x_t, m_t) \end{bmatrix}}_{\text{Hamiltonian}} - \underbrace{\begin{bmatrix} 0 \\ (\gamma + \lambda)(m_t - \nabla \mathcal{K}^*(\lambda x_t)) \end{bmatrix}}_{\text{Descent}},$$

where the Hamiltonian component is orthogonal to the gradient $[\nabla_x H, \nabla_m H]$ of $H(x, m)$ and preserves the total energy $H(x, m)$, and the descent component introduces a damping like effect to decrease the energy $H(x, m)$. $\square$

B.6 LION-$\mathcal{K}$ WITHOUT WEIGHT DECAY – A HAMILTONIAN + DESCENT DERIVATION

When the weight decay in Lion-$\mathcal{K}$ is turned off ($\lambda = 0$), there is an alternative way to analyze it that is amendable to the Hamiltonian + descent structure in (12).

Recall that the Lion-$\mathcal{K}$ ODE is of the following form when $\lambda = 0$:

$$\begin{aligned} &\dot{x}_t = \nabla\mathcal{K}(\tilde{m}_t), \qquad \tilde{m}_t = m_t - \varepsilon(\alpha\nabla f(x_t) + \gamma m_t) \\ &\dot{m}_t = -\alpha\nabla f(x_t) - \gamma m_t \end{aligned} \tag{29}$$

Assume $\varepsilon\gamma < 1$. Define $\tilde{\mathcal{K}}(m) = \frac{1}{1-\varepsilon\gamma}\mathcal{K}((1-\varepsilon\gamma)m)$, and the following Lyapunov function:

$$H(x,m) = \alpha f(x) + \tilde{\mathcal{K}}(m) = \alpha f(x) + \frac{1}{1-\varepsilon\gamma}\mathcal{K}((1-\varepsilon\gamma)m). \tag{30}$$

Note that $\nabla_x H(x,m) = \alpha\nabla f(x)$ and $\nabla_m H(x,m) = \nabla\mathcal{K}((1-\varepsilon)m)$. One can decompose (29) into the following Hamiltonian + descent decomposition:

$$\begin{bmatrix}\dot{x}_t \\ \dot{m}_t\end{bmatrix} = \underbrace{\begin{bmatrix} +\nabla_m H(x_t,m_t) \\ -\nabla_x H(x_t,m_t)\end{bmatrix}}_{\text{Hamiltonian}} - \underbrace{\begin{bmatrix}\nabla\mathcal{K}(\tilde{m}_t^0) - \nabla\mathcal{K}(\tilde{m}_t) \\ \gamma m_t\end{bmatrix}}_{\text{Descent}},$$

where we define $\tilde{m}_t^0 = (1-\varepsilon\gamma)m_t$ and hence $\tilde{m}_t - \tilde{m}_t^0 = -\varepsilon\alpha\nabla f(x_t)$.

Using the monotonicity of subgradient (Lemma 2.1), one can show that the second component in the decomposition above is a descent direction of $H(x,m)$ in (30):

1) Let $\hat{\nabla}_x H_t \coloneqq -\nabla\mathcal{K}(\tilde{m}_t^0) + \nabla\mathcal{K}(\tilde{m}_t)$, then it is a descent direction of $H(x,m)$, because

$$\begin{aligned} \nabla_x H(x_t,m_t)^\top \hat{\nabla}_x H_t &= \alpha\nabla f(x_t)^\top \hat{\nabla}_x H_t \\ &= -\frac{1}{\varepsilon}(\tilde{m}_t^0 - \tilde{m}_t)^\top(\nabla\mathcal{K}(\tilde{m}_t^0) - \nabla\mathcal{K}(\tilde{m}_t)) \le 0, \end{aligned}$$

where we used the monotonicity of $\nabla\mathcal{K}(\cdot)$.

2) If $m = 0$ is the minimum of $\mathcal{K}$, then $\hat{\nabla}_m H_t \coloneqq -\gamma m_t$ is a descent direction of $H(x,m)$ because,

$$\nabla_m H(x_t,m_t)^\top \hat{\nabla}_m H_t = -\gamma\nabla\mathcal{K}((1-\varepsilon\gamma)m_t)^\top m_t \le \frac{\gamma}{1-\varepsilon\gamma}(\mathcal{K}(0) - \mathcal{K}((1-\varepsilon\gamma)m_t)) \le 0.$$

Hence, we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}H(x_t,m_t) &= \nabla_x H(x_t,m_t)^\top \hat{\nabla}_x H_t + \nabla_m H(x_t,m_t)^\top \hat{\nabla}_m H_t \\ &= -\frac{1}{\varepsilon}(\tilde{m}_t^0 - \tilde{m}_t)^\top(\nabla\mathcal{K}(\tilde{m}_t^0) - \nabla\mathcal{K}(\tilde{m}_t)) - \gamma\nabla\mathcal{K}((1-\varepsilon\gamma)m_t)^\top m_t \le 0. \end{aligned}$$

Moreover, if $m = 0$ is the unique minimum of $\mathcal{K}$, and $\varepsilon\gamma < 1$, then $\nabla\mathcal{K}((1-\varepsilon\gamma)m_t)^\top m_t = 0$ implies that $m_t = 0$, and one can show that the equilibrium points of (29) are stationary points of $H(x,m)$ using LaSalle's invariance principle.

B.7 MAIN RESULT OF LION-$\mathcal{K}$ ODE

Theorem B.3. *Assume $\mathcal{K}$ is convex with conjugate $\mathcal{K}^*$. Assume $f, \mathcal{K}, \mathcal{K}^*$ are continuously differentiable. Assume (x_t, m_t) is the solution of the following ODE:*

$$\begin{aligned} &\dot{x}_t = \nabla\mathcal{K}(\tilde{m}_t) - \lambda x_t, \qquad \textit{with} \qquad \tilde{m}_t = m_t - \varepsilon(\gamma m_t + \alpha\nabla f(x_t)), \\ &\dot{m}_t = -\alpha\nabla f(x_t) - \gamma m_t, \end{aligned}$$

where $\alpha, \gamma, \lambda, \varepsilon > 0$ and $\epsilon\gamma \le 1$. Let

$$H(x,m) = \alpha f(x) + \frac{\gamma}{\lambda}\mathcal{K}^*(\lambda x) + \frac{1-\varepsilon\gamma}{1+\varepsilon\lambda}(\mathcal{K}^*(\lambda x) + \mathcal{K}(m) - \lambda m^\top x).$$

Then H yields a Lyapunov function in that

$$-\frac{\mathrm{d}}{\mathrm{d}t}H(x_t, m_t) = \Delta(x_t, m_t) \coloneqq \frac{\lambda+\gamma}{1+\varepsilon\lambda}\Delta_1(x_t, \tilde{m}_t) + \frac{1-\varepsilon\gamma}{(1+\varepsilon\lambda)}\Delta_2(m_t, \tilde{m}_t) \geq 0,$$

where

$$\Delta_1(x, \tilde{m}) = (\tilde{m} - \nabla\mathcal{K}^*(\lambda x))^\top(\nabla\mathcal{K}(\tilde{m}) - \lambda x),$$
$$\Delta_2(m, \tilde{m}) = \frac{1}{\varepsilon}(\tilde{m} - m)^\top(\nabla\mathcal{K}(\tilde{m}) - \nabla\mathcal{K}(m)).$$

Moreover, the accumulation points of all trajectories are stationary points of $F(x) = \alpha f(x) + \frac{\gamma}{\lambda}\mathcal{K}^*(\lambda x)$.

Proof. It is not obvious how to construct the Lyapunov function directly from the ODE. The following proof describes the process of discovering $H(x, m)$. We start by examing what inequalities we can write down using the monotonicity of $\nabla\mathcal{K}$ and $\nabla\mathcal{K}^*$ via Lemma 2.1, and then work out the Lyapunov function backward.

Write $\tilde{m} = m - \varepsilon(\gamma m + \alpha\nabla f(x))$. Because $\nabla\mathcal{K}$ is a monotonic mapping, we have by Lemma 2.1 the following key inequalities:

$$(-\tilde{m} + \nabla\mathcal{K}^*(\lambda x))^\top(\nabla\mathcal{K}(\tilde{m}) - \lambda x) \leq 0,$$
$$(m - \tilde{m})^\top(\nabla\mathcal{K}(\tilde{m}) - \nabla\mathcal{K}(m)) \leq 0,$$

or equivalently

$$(\varepsilon\alpha\nabla f(x) - (1-\varepsilon\gamma)m + \nabla\mathcal{K}^*(\lambda x))^\top(\nabla\mathcal{K}(\tilde{m}) - \lambda x) \leq 0 \tag{31}$$
$$\varepsilon(\alpha\nabla f(x) + \gamma m)^\top((\nabla\mathcal{K}(\tilde{m}) - \lambda x) - (\nabla\mathcal{K}(m) - \lambda x)) \leq 0 \tag{32}$$

Write $V_x = \nabla\mathcal{K}(\tilde{m}) - \lambda x$, and $V_m = -\alpha\nabla f(x) - \gamma m$. So the ODE is $\dot{x} = V_x$ and $\dot{m} = V_m$. The inequalities can be rewritten into

$$(\varepsilon\alpha\nabla f(x) - (1-\varepsilon\gamma)m + \nabla\mathcal{K}^*(\lambda x))^\top V_x \leq 0 \tag{33}$$
$$-\varepsilon V_m^\top(V_x - (\nabla\mathcal{K}(m) - \lambda x)) \leq 0 \tag{34}$$

Taking $\frac{1}{\varepsilon(1+\eta)}(Eq.\ (33) + \eta \times Eq.\ (34))$ for any $\eta \geq 0$, we get

$$\left(\alpha\nabla f(x) - \frac{1-\varepsilon\gamma(1+\eta)}{\varepsilon(1+\eta)}m + \frac{1}{\varepsilon(1+\eta)}\nabla\mathcal{K}^*(\lambda x)\right)^\top V_x + \frac{\eta\varepsilon}{\varepsilon(1+\eta)}(\nabla\mathcal{K}(m) - \lambda x)^\top V_m \leq 0$$

Define

$$\tilde{H}(x, m) = \alpha f(x) + \frac{1}{\varepsilon(1+\eta)\lambda}\mathcal{K}^*(\lambda x) + \frac{1-\varepsilon\gamma(1+\eta)}{\varepsilon(1+\eta)}\frac{1}{\lambda}\mathcal{K}(m) - \frac{1-\varepsilon\gamma(1+\eta)}{\varepsilon(1+\eta)}m^\top x.$$

Then the inequality was reduced to

$$\nabla_x\tilde{H}(x, m)^\top V_x + \frac{\varepsilon\eta\lambda}{1-\varepsilon\gamma(1+\eta)}\nabla_m\tilde{H}(x, m)^\top V_m \leq 0.$$

If we take η such that

$$\frac{\varepsilon\eta\lambda}{1-\varepsilon\gamma(1+\eta)} = 1, \tag{35}$$

then we have when following $\dot{x} = V_x$ and $\dot{m} = V_m$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{H}(x, m) = \nabla_x\tilde{H}(x, m)^\top V_x + \nabla_m\tilde{H}(x, m)^\top V_m \leq 0.$$

Furthermore, when (35) holds, we have

$$\eta = \frac{1-\varepsilon\gamma}{\varepsilon(\lambda+\gamma)}, \qquad \frac{1}{\varepsilon(1+\eta)} = \frac{\lambda+\gamma}{1+\varepsilon\lambda}, \qquad \frac{1-\varepsilon\gamma(1+\eta)}{\varepsilon(1+\eta)} = \frac{(1-\varepsilon\gamma)\lambda}{1+\varepsilon\lambda}, \tag{36}$$

and hence

$$\begin{aligned}\tilde{H}(x,m) &= \alpha f(x) + \left(\frac{\lambda+\gamma}{(1+\epsilon\lambda)\lambda} - \frac{1-\varepsilon\gamma}{1+\varepsilon\lambda}\right)\mathcal{K}^*(\lambda x) + \frac{1-\varepsilon\gamma}{1+\varepsilon\lambda}\left(\mathcal{K}^*(\lambda x) + \mathcal{K}(m) - \lambda m^\top x\right) \\ &= \alpha f(x) + \frac{\gamma}{\lambda}\mathcal{K}^*(\lambda x) + \frac{1-\varepsilon\gamma}{1+\varepsilon\lambda}\left(\mathcal{K}^*(\lambda x) + \mathcal{K}(m) - \lambda m^\top x\right) \\ &= H(x,m).\end{aligned}$$

In this case,

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}H(x,m) \\ &= \frac{1}{\varepsilon(1+\eta)}(Eq.~(33) + \eta \times Eq.~(34)) \\ &= \frac{\lambda+\gamma}{1+\varepsilon\lambda} \times Eq.~(33) + \frac{1-\varepsilon\gamma}{\varepsilon(1+\varepsilon\lambda)} \times Eq.~(34). \qquad // \frac{\eta}{\varepsilon(1+\eta)} = \frac{1-\varepsilon\gamma}{\varepsilon(1+\varepsilon\lambda)} \text{ from } (36) \\ &= -\frac{\lambda+\gamma}{1+\varepsilon\lambda}(\tilde{m} - \nabla\mathcal{K}^*(\lambda x))^\top(\nabla\mathcal{K}(\tilde{m}) - \lambda x) - \frac{1-\epsilon\gamma}{(1+\epsilon\lambda)\varepsilon}(\tilde{m} - m)^\top(\nabla\mathcal{K}(\tilde{m}) - \nabla\mathcal{K}(m)) \le 0.\end{aligned}$$

To ensure that $\eta \ge 0$, we need $\varepsilon\gamma \le 1$. □

LaSalle's invariance principle Let $H(z)$ is a continuously differentiable Lyapunov function of $\frac{\mathrm{d}}{\mathrm{d}t}z_t = v(z_t)$, satisfying $\frac{\mathrm{d}}{\mathrm{d}t}H(z_t) \le 0$. By LaSalle's Invariance Principle, the accumulation points of any trajectories of $\frac{\mathrm{d}}{\mathrm{d}t}z_t = v(z_t)$ is included in

$$\mathcal{I} = \{\text{the union of all trajectories } z_t \text{ satisfying } \frac{\mathrm{d}}{\mathrm{d}t}H(z_t) = 0 \text{ for all } t \ge 0\}.$$

For the Lion-$\mathcal{K}$ ODE and its H, the points in $\mathcal{I}$ should satisfy $\tilde{m}_t = \nabla\mathcal{K}^*(\lambda x_t)$, which yields $\nabla\mathcal{K}(\tilde{m}_t) = \lambda x_t$, and hence

$$\dot{x}_t = \nabla\mathcal{K}(\tilde{m}_t) - \lambda x_t = 0.$$

This suggests that x_t is constant for the trajectories in $\mathcal{I}$. Because $\tilde{m}_t = \nabla\mathcal{K}^*(\lambda x_t)$ and $\tilde{m}_t = m_t - \varepsilon(\alpha\nabla f(x_t) + \gamma m_t)$, we have

$$(1-\varepsilon\gamma)m_t = \nabla\mathcal{K}^*(\lambda x_t) + \varepsilon\alpha\nabla f(x_t)$$

Hence, $(1-\varepsilon\gamma)m_t$ is also constants in the trajectories in $\mathcal{I}$. This suggests that $(1-\varepsilon\gamma)\dot{m}_t = 0$ along the trajectories in $\mathcal{I}$, and hence

$$\begin{aligned}0 &= (1-\varepsilon\gamma)\dot{m}_t \\ &= -(1-\varepsilon\gamma)(\alpha\nabla f(x_t) + \gamma m_t) \\ &= -(1-\varepsilon\gamma)\alpha\nabla f(x_t) - \gamma\nabla\mathcal{K}^*(\lambda x_t) - \varepsilon\gamma\alpha\nabla f(x_t) \\ &= -\alpha\nabla f(x_t) - \gamma\nabla\mathcal{K}^*(\lambda x_t) \\ &= -\nabla F(x_t) \qquad // F(x) = \alpha f(x) + \frac{\gamma}{\lambda}\mathcal{K}^*(\lambda x)\end{aligned}$$

Hence, all trajectories in $\mathcal{I}$ are singleton points and are stationary points of the objective $F(x) = \alpha f(x) + \frac{\gamma}{\lambda}\mathcal{K}^*(\lambda x)$.

B.8 The Decomposition Structure

We provide the decomposition structure (11) which provides a simplified proof of the Lyapunov property.

Lemma B.4. *For ODE $\dot{x}_t = V_x(x_t, m_t)$, $\dot{m}_t = V_m(x_t, m_t)$, let $H(x,m)$ be a function satisfying*

$$\nabla_x H(x,m) = -\tilde{V}_x(x,m) + \eta V_m(x,m)$$
$$\nabla_m H(x,m) = -\hat{V}_m(x,m) - \eta V_x(x,m),$$

where $a \in \mathbb{R}$ and $\hat{V}_x$ and $\hat{V}_m$ have positive inner products with V_x, V_m, respectively, that is,

$$\hat{V}_x(x,m)^\top V_x(x,m) \geq 0, \qquad \hat{V}_m(x,m)^\top V_m(x,m) \geq 0, \quad \forall x, m.$$

Then we have

$$\frac{\mathrm{d}}{\mathrm{d}t} H(x_t, m_t) \leq 0.$$

Proof.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} H(x_t, m_t) &= \nabla_x H^\top V_x + \nabla_m H^\top V_m \\ &= (-\hat{V}_x + aV_m)^\top V_x + (-\hat{V}_m - aV_x)^\top V_m \\ &= -(\hat{V}_x^\top V_x + \hat{V}_m^\top V_m) \leq 0.\end{aligned}$$

□

Lemma B.5. *Under the condition of Theorem 3.1, let*

$$\begin{aligned}V_x(x,m) &= \nabla \mathcal{K}(\tilde{m}) - \lambda x \\ V_m(x,m) &= -\alpha \nabla f(x) - \gamma m = \frac{\tilde{m} - m}{\varepsilon}\end{aligned}$$

and related

$$\begin{aligned}\hat{V}_x(x,m) &= \tilde{m} - \nabla \mathcal{K}^*(\lambda x) = -\varepsilon\alpha \nabla f(x) + (1 - \varepsilon\gamma)m - \nabla \mathcal{K}^*(\lambda x), \\ \hat{V}_m(x,m) &= \nabla \mathcal{K}(\tilde{m}) - \nabla \mathcal{K}(m).\end{aligned}$$

Then we have $\hat{V}_x^\top V_x \geq 0$ and $\hat{V}_m^\top V_m \geq 0$ by Lemma 2.1. Moreover,

$$\begin{aligned}\nabla_x H(x,m) &= -\eta' \hat{V}_x - \eta V_m \\ \nabla_m H(x,m) &= -\eta \hat{V}_m + \eta V_x,\end{aligned}$$

where $\eta = \frac{1-\varepsilon\gamma}{1+\varepsilon\lambda}$ and $\eta' = \frac{\gamma+\lambda}{1+\varepsilon\lambda}$. This yields

$$\frac{\mathrm{d}}{\mathrm{d}t} H(x_t, m_t) = \nabla_x H^\top V_x + \nabla_m H^\top V_m = -(\eta' \hat{V}_x^\top V_x + \eta \hat{V}_m^\top V_m) \leq 0.$$

Proof. Let $\eta = \frac{1-\varepsilon\gamma}{1+\varepsilon\lambda}$. We have We have

$$\begin{aligned}\nabla_m H(x,m) &= \eta(\nabla \mathcal{K}(m) - \lambda x) \\ &= \eta(\nabla \mathcal{K}(\tilde{m}) - \lambda x + \nabla \mathcal{K}(m) - \nabla \mathcal{K}(\tilde{m})) \\ &= \eta(V_x - \hat{V}_m).\end{aligned}$$

$$\begin{aligned}&\nabla_x H(x,m) \\ &= \alpha \nabla f(x) + \gamma \nabla \mathcal{K}^*(\lambda x) + \eta(\lambda \nabla \mathcal{K}^*(\lambda x) - \lambda m) \\ &= \alpha \nabla f(x) + (\gamma + \eta\lambda) \nabla \mathcal{K}^*(\lambda x) - \eta\lambda m \\ &= (\gamma + \eta\lambda)(\varepsilon\alpha \nabla f(x) - (1 - \varepsilon\gamma)m + \nabla \mathcal{K}^*(\lambda x)) + (\alpha - (\gamma + \eta\lambda)\varepsilon\alpha) \nabla f(x) - (\eta\lambda - (\gamma + \eta\lambda)(1 - \varepsilon\gamma))m \\ &= \frac{\gamma + \lambda}{1 + \varepsilon\lambda}(\varepsilon\alpha \nabla f(x) - (1 - \varepsilon\gamma)m + \nabla \mathcal{K}^*(\lambda x)) + \eta\alpha \nabla f(x) + \eta\gamma m \\ &= -\frac{\gamma + \lambda}{1 + \varepsilon\lambda} \hat{V}_x - \eta V_m,\end{aligned}$$

where we used the following identities on η:

$$\begin{aligned}&(\gamma + \eta\lambda) = \gamma + \frac{1 - \varepsilon\gamma}{1 + \varepsilon\lambda}\lambda = \frac{\gamma + \lambda}{1 + \varepsilon\lambda} \\ &1 - (\gamma + \eta\lambda)\varepsilon = 1 - \frac{\gamma + \lambda}{1 + \varepsilon\lambda}\varepsilon = \frac{1 - \varepsilon\gamma}{1 + \varepsilon\lambda} = \eta \\ &\eta\lambda - (\gamma + \eta\lambda)(1 - \varepsilon\gamma) = -\gamma + \frac{\gamma + \lambda}{1 + \varepsilon\lambda}\varepsilon\gamma = \frac{\varepsilon\gamma^2 - \gamma}{1 + \varepsilon\lambda} = -\gamma\eta.\end{aligned}$$

□

B.9 Constraint Enforcing: Continuous Time

When $\mathcal{K}^*$ can possible take infinite values, the minimization of $H(x, m)$ becomes a constrained optimization. Let $\mathrm{dom}\mathcal{K}^* = \{x \colon \mathcal{K}^*(x) < +\infty\}$. The optimization can be framed as

$$\min_{x,m} H(x, m) \quad s.t. \quad \lambda x \in \mathrm{dom}\mathcal{K}^*.$$

The Lion-$\mathcal{K}$ algorithm would first steer x_t to the region where $\mathcal{K}^*$ has finite values, and then decrease the finite parts of the objective function. In the following, we show that Lion-$\mathcal{K}$ enforces the constraint with a fast linear rate: the distance from λx_t and $\mathrm{dom}\mathcal{K}^*$ decays exponentially fast with time t, and once $\lambda x_{t_0} \in \mathrm{dom}\mathcal{K}^*$, then λx_t stays within $\mathrm{dom}\mathcal{K}^*$ for all $t > t_0$.

Theorem B.6. *Under the condition of Theorem 3.1, we have*

$$\mathrm{dist}(\lambda x_t, \mathrm{dom}\mathcal{K}^*) \leq \exp(\lambda(s - t))\, \mathrm{dist}(\lambda x_s, \mathrm{dom}\mathcal{K}^*).$$

Proof. Define $w_{s\to t} = \exp(\lambda(s - t))$. Integrating $\dot{x}_t = \nabla\mathcal{K}(\tilde{m}_t) - \lambda x_t$, we have

$$\lambda x_t = (1 - w_{s\to t})z_{s\to t} + w_{s\to t}(\lambda x_s), \quad \text{where} \quad z_{s\to t} = \frac{\int_s^t w_{\tau\to t}\nabla\mathcal{K}(\tilde{m}_\tau)\mathrm{d}s}{\int_s^t w_{\tau\to t} d\tau}, \quad \forall 0 \leq s \leq t.$$

We have $\nabla\mathcal{K}(\tilde{m}_\tau) \in \mathrm{dom}\mathcal{K}^*$ from Lemma B.7 and $\mathrm{dom}\mathcal{K}^*$ is convex. Hence $z_{s\to t}$, as the convex combination of $\{\nabla\mathcal{K}(\tilde{m}_\tau\}_\tau$, belongs to $\mathrm{dom}\mathcal{K}^*$. For any $\epsilon > 0$, let $\lambda\hat{x}_s \in \mathrm{dom}\mathcal{K}^*$ to the point satisfying $\|\lambda\hat{x}_s - \lambda x_s\| \leq \mathrm{dist}(\lambda x_s, \mathrm{dom}\mathcal{K}^*) + \epsilon$. Hence,

$$\begin{aligned}
\mathrm{dist}(\lambda x_t,\ \mathrm{dom}\mathcal{K}^*) &= \inf_{z\in\mathrm{dom}\mathcal{K}^*} \|\lambda x_t - z\| \\
&\leq \|\lambda x_t - (1 - w_{s\to t})z_{s\to t} - w_{s\to t}\lambda\hat{x}_s)\| \\
&= w_{s\to t}\, \|\lambda x_s - \lambda\hat{x}_s\| \\
&\leq \exp(\lambda(s - t))(\mathrm{dist}(\lambda x_s, \mathrm{dom}\mathcal{K}^*) + \epsilon).
\end{aligned}$$

Taking $\epsilon \to 0$ yields

$$\mathrm{dist}(\lambda x_t, \mathrm{dom}\mathcal{K}^*) \leq \exp(\lambda(s - t))\, \mathrm{dist}(\lambda x_s, \mathrm{dom}\mathcal{K}^*).$$

□

Lemma B.7. *Assume $\mathcal{K}$ is proper, closed and convex, and $\mathcal{K}^*$ is the conjugate of $\mathcal{K}$. We have*

$$\partial\mathcal{K}(z) \subseteq \mathrm{dom}\mathcal{K}^*, \quad \forall z \in \mathrm{dom}\mathcal{K}.$$

Proof. If $x \in \partial\mathcal{K}(z)$, then z attains the minimum of $\mathcal{K}^*(x) = \sup_z\{x^\top z - \mathcal{K}(z)\}$, suggesting that $\mathcal{K}^*(x) = x^\top z - \mathcal{K}(z) < +\infty$, and hence $x \in \mathrm{dom}\mathcal{K}^*$. □

B.10 Discrete Time Analysis

Theorem B.8. *Assume $f \colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth, and $\mathcal{K} \colon \mathbb{R}^d \to \mathbb{R}$ is closed and convex. Consider the following scheme:*

$$\begin{aligned}
m_{t+1} &= \beta_2 m_t - (1 - \beta_2)\nabla f(x_t) \\
\tilde{m}_{t+1} &= \beta_1 m_t - (1 - \beta_1)\nabla f(x_t) \\
x_{t+1} &= x_t + \epsilon(\nabla\mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1}),
\end{aligned} \tag{37}$$

where $\nabla\mathcal{K}$ is a subgradient of $\mathcal{K}$, and $\beta_1, \beta_2 \in (0, 1)$, and $\beta_2 > \beta_1$, and $\epsilon, \lambda > 0$. Let $\mathcal{K}^*$ be the conjugate function of $\mathcal{K}$. Define the following Lyapunov function:

$$H(x, m) = f(x) + \frac{1}{\lambda}\mathcal{K}^*(\lambda x) + \frac{\beta_1}{\epsilon\lambda(1 - \beta_1) + (1 - \beta_2)}(\mathcal{K}^*(\lambda x) + \mathcal{K}(m) - \lambda x^\top m),$$

and

$$\begin{aligned}
\Delta_t^1 &= (\nabla\mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top(\tilde{m}_{t+1} - \nabla\mathcal{K}^*(\lambda x_{t+1})), \\
\Delta_t^2 &= (\nabla\mathcal{K}(\tilde{m}_{t+1}) - \nabla\mathcal{K}(m_{t+1}))^\top(\tilde{m}_{t+1} - m_{t+1}),
\end{aligned}$$

where $\nabla \mathcal{K}^*$ is a subgradient of $\mathcal{K}^*$. Then we have $\Delta_t^1 \geq 0$ and $\Delta_t^2 \geq 0$ from Lemma B.9, and

$$H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \leq -\epsilon(a\Delta_t^1 + b\Delta_t^2) + \frac{L\epsilon^2}{2} \|\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1}\|_2^2,$$

where

$$a = \frac{\epsilon\lambda\beta_1}{\epsilon\lambda(1-\beta_1) + (1-\beta_2)} + 1 \geq 0, \qquad b = \frac{\beta_1(1-\beta_2)}{(\beta_2 - \beta_1)(\epsilon\lambda(1-\beta_1) + (1-\beta_2))} \geq 0.$$

Hence, a telescoping sum yields

$$\frac{1}{T}\sum_{t=0}^{T-1} a\Delta_t^1 + b\Delta_t^2 \leq \frac{H(x_0, m_0) - H(x_T, m_T)}{\epsilon T} + \frac{L\epsilon}{2} B_t,$$

where $B_t = \frac{1}{T}\sum_{t=1}^{T} \|\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1}\|_2^2$.

Note that we used an implicit scheme in the update of x_t in (43). It is equivalent the explicit scheme with an adjusted learning rate:

$$x_{t+1} = x_t + \frac{\epsilon}{1+\epsilon\lambda}(\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_t).$$

Proof. We follow the proof in the continuous-time case to find out a Lyapunov function for the discrete time update in (43). We start with constructing the basic inequalities and work out the Lyapunov function backwardly. From Lemma 2.1, we have

$$(\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top (\nabla \mathcal{K}^*(\lambda x_{t+1}) - \tilde{m}_{t+1}) \leq 0. \tag{38}$$

$$(\nabla \mathcal{K}(\tilde{m}_{t+1}) - \nabla \mathcal{K}(m_{t+1}))^\top (m_{t+1} - \tilde{m}_{t+1}) \leq 0. \tag{39}$$

Taking $a \times Eq.(38) + b \times Eq(39)$ for $a, b \geq 0$, we have

$$(\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top (a(\nabla \mathcal{K}^*(\lambda x_{t+1}) - \tilde{m}_{t+1}) + b(m_{t+1} - \tilde{m}_{t+1})) + \cdots \\ + b(\nabla \mathcal{K}(m_{t+1}) - \lambda x_{t+1})^\top (-m_{t+1} + \tilde{m}_{t+1}) \leq 0.$$

Plugging (43) yields

$$(\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top (a\nabla \mathcal{K}^*(\lambda x_{t+1}) - ((a+b)\beta_1 - b\beta_2)m_t + (a - (a+b)\beta_1 + b\beta_2)\nabla f(x_t)) \\ - b(\beta_2 - \beta_1)(\nabla \mathcal{K}(m_{t+1}) - \lambda x_{t+1})^\top (m_t + \nabla f(x_t)) \leq 0$$

Define

$$H(x, m) = (a-c)f(x) + \frac{a}{\lambda}\mathcal{K}^*(\lambda x) + \frac{c}{\lambda}\mathcal{K}(m) - cx^\top m, \quad \text{with} \quad c = (a+b)\beta_1 - b\beta_2,$$

and

$$\hat{\nabla}_x H_t = (a-c)\nabla f(x_t) + a\nabla \mathcal{K}^*(\lambda x_{t+1}) - cm_t, \qquad \hat{\nabla}_m H_t = \frac{c}{\lambda}\nabla \mathcal{K}(m_{t+1}) - cx_{t+1}.$$

Then the inequality can be written into

$$\hat{\nabla}_x H_t^\top (\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1}) + \hat{\nabla}_m H_t^\top \left(\frac{b(\beta_2 - \beta_1)\lambda}{c}(-m_t - \nabla f(x_t))\right) \leq 0.$$

Plugging the update rule of $x_{t+1} = x_t + \epsilon(\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})$ and $m_{t+1} - m_t = -(1-\beta_2)(m_t + \nabla f(x_t))$, we get

$$\hat{\nabla}_x H_t^\top \left(\frac{x_{t+1} - x_t}{\epsilon}\right) + \hat{\nabla}_m H_t^\top \left(\frac{b(\beta_2 - \beta_1)\lambda}{c(1-\beta_2)}(m_{t+1} - m_t)\right) \leq 0.$$

To make this coincide with the linear approximation of the difference $H(x_{t+1}, m_{t+1}) - H(x_t, m_t)$ (see Lemma B.9), we want

$$\frac{b(\beta_2 - \beta_1)\lambda}{c(1-\beta_2)} = \frac{1}{\epsilon}.$$

On the other hand, to make the coefficient of $f(x)$ in $H(x,m)$ equal to one, we want $a - c = 1$. This yields the following equations on a, b, c:

$$c = (a+b)\beta_1 - b\beta_2, \qquad \frac{b(\beta_2 - \beta_1)\lambda}{c(1-\beta_2)} = \frac{1}{\epsilon}, \qquad a - c = 1, \qquad a, b \geq 0.$$

To solve this, let $c = z\epsilon(\beta_2 - \beta_1)\lambda$ and $b = z(1-\beta_2)$ for some $z \geq 0$ and plug them together with $a = c + 1$ into the first equations:

$$z\epsilon(\beta_2 - \beta_1)\lambda = (z\epsilon(\beta_2 - \beta_1)\lambda + 1 + z(1-\beta_2))\beta_1 - z(1-\beta_2)\beta_2.$$

We get

$$\begin{aligned} z &= \frac{\beta_1}{\epsilon(\beta_2 - \beta_1)\lambda - \epsilon(\beta_2 - \beta_1)\lambda\beta_1 - (1-\beta_2)\beta_1 + (1-\beta_2)\beta_2} \\ &= \frac{\beta_1}{\epsilon\lambda(\beta_2 - \beta_1)(1-\beta_1) + (1-\beta_2)(\beta_2 - \beta_1)} \\ &= \frac{\beta_1}{(\beta_2 - \beta_1)(\epsilon\lambda(1-\beta_1) + (1-\beta_2))} \geq 0. \end{aligned}$$

Hence

$$b = \frac{\beta_1(1-\beta_2)}{(\beta_2 - \beta_1)(\epsilon\lambda(1-\beta_1) + (1-\beta_2))} \geq 0, \quad c = \frac{\epsilon\lambda\beta_1}{\epsilon\lambda(1-\beta_1) + (1-\beta_2)} \geq 0, \quad a = c + 1 \geq 0.$$

In this case, we have

$$\begin{aligned} H(x,m) &= f(x) + \frac{1}{\lambda}\mathcal{K}^*(\lambda x) + c(\mathcal{K}^*(\lambda x) + \mathcal{K}(m) - \lambda x^\top m) \\ &= f(x) + \frac{1}{\lambda}\mathcal{K}^*(\lambda x) + \frac{\epsilon\lambda\beta_1}{\epsilon\lambda(1-\beta_1) + (1-\beta_2)}(\mathcal{K}^*(\lambda x) + \mathcal{K}(m) - \lambda x^\top m), \end{aligned}$$

and

$$\hat{\nabla}_x H_t^\top \left(\frac{x_{t+1} - x_t}{\epsilon}\right) + \hat{\nabla}_m H_t^\top \left(\frac{m_{t+1} - m_t}{\epsilon}\right) = -a\Delta_t^1 - b\Delta_t^2 \leq 0.$$

From Lemma B.9, we get

$$H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \leq -\epsilon(a\Delta_t^1 + b\Delta_t^2) + \frac{L}{2}\left\|x_{t+1} - x_t\right\|_2^2.$$

□

Lemma B.9. *Let $H(x,m) = f(x) + \mathcal{K}_1(x) + \mathcal{K}_2(m) - \lambda xm$, where f is L-smooth, and $\mathcal{K}_1, \mathcal{K}_2$ are convex functions with subgradient $\nabla\mathcal{K}_1$ and $\nabla\mathcal{K}_2$. Then*

$$H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \leq \hat{\nabla}_x H_t^\top(x_{t+1} - x_t) + \hat{\nabla}_m H_t^\top(m_{t+1} - m_t) + \frac{L}{2}\left\|x_{t+1} - x_t\right\|_2^2,$$

where

$$\begin{aligned} \hat{\nabla}_x H_t &= \nabla f(x_t) + \mathcal{K}_1(x_{t+1}) - \lambda m_t \\ \hat{\nabla}_m H_t &= \mathcal{K}_2(m_{t+1}) - \lambda x_{t+1}. \end{aligned}$$

Note the use of x_t vs. x_{t+1} and m_t vs. m_{t+1} in $\hat{\nabla}_x H_t$ and $\hat{\nabla}_m H_t$.

Proof. We have

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq \nabla f(x_t)^\top(x_{t+1} - x_t) + \frac{L}{2}\left\|x_{t+1} - x_t\right\|_2^2 \\ \mathcal{K}_1(x_{t+1}) - \mathcal{K}_1(x_t) &\leq \nabla\mathcal{K}_1(x_{t+1})^\top(x_{t+1} - x_t) \\ \mathcal{K}_2(m_{t+1}) - \mathcal{K}_2(m_t) &\leq \nabla\mathcal{K}_2(m_{t+1})^\top(m_{t+1} - m_t) \\ x_{t+1}^\top m_{t+1} - x_t^\top m_t &= m_t^\top(x_{t+1} - x_t) + x_{t+1}^\top(m_{t+1} - m_t). \end{aligned}$$

Summing them together yields the result. □

Theorem B.10. *Under the same conditions of Theorem 4.1, for any two integers $s \leq t$,*

$$\mathrm{dist}(\lambda x_t, \mathrm{dom}\mathcal{K}^*) \leq \left(\frac{1}{1+\epsilon\lambda}\right)^{s-t} \mathrm{dist}(\lambda x_s, \mathrm{dom}\mathcal{K}^*), \quad \forall s \leq t.$$

Proof. Rewriting the update into the explicit form:

$$x_{t+1} = \frac{1}{1+\epsilon\lambda} x_t + \frac{\epsilon}{1+\epsilon\lambda} \nabla\mathcal{K}(\tilde{m}_{t+1}).$$

Unrolling this update yields, with $w_{s\to t} = \left(\frac{1}{1+\epsilon\lambda}\right)^{s-t}$,

$$\lambda x_t = (1 - w_{s\to t}) z_{s\to t} + w_{s\to t} x_s, \qquad z_{s\to t} = \frac{\sum_{k=s+1}^{t} w_{k\to t} \nabla\mathcal{K}(\tilde{m}_k)}{\sum_{k=s+1}^{t} w_{k\to t}}.$$

We have $\nabla\mathcal{K}(\tilde{m}_k) \in \mathrm{dom}\mathcal{K}^*$ from Lemma B.7 and $\mathrm{dom}\mathcal{K}^*$ is convex. Hence $z_{s\to t}$, as the convex combination of $\{\nabla\mathcal{K}(\tilde{m}_k\}_k$, belongs to $\mathrm{dom}\mathcal{K}^*$. For any $\epsilon > 0$, let $\lambda\hat{x}_s \in \mathrm{dom}\mathcal{K}^*$ to the point satisfying $\|\lambda\hat{x}_s - \lambda x_s\| \leq \mathrm{dist}(\lambda x_s, \mathrm{dom}\mathcal{K}^*) + \epsilon$. Hence,

$$\begin{aligned}
\mathrm{dist}(\lambda x_t,\ \mathrm{dom}\mathcal{K}^*) &= \inf_{z\in\mathrm{dom}\mathcal{K}^*} \|\lambda x_t - z\| \\
&\leq \|\lambda x_t - (1 - w_{s\to t}) z_{s\to t} + w_{s\to t}\lambda\hat{x}_s)\| \\
&= w_{s\to t} \|\lambda x_s - \lambda\hat{x}_s\| \\
&\leq \left(\frac{1}{1+\epsilon\lambda}\right)^{s-t} (\mathrm{dist}(\lambda x_s, \mathrm{dom}\mathcal{K}^*) + \epsilon).
\end{aligned}$$

Taking $\epsilon \to 0$ yields the result.

□

B.11 Analysis with Stochastic Gradient for Lion-$\mathcal{K}$

In this section, we are going to have the convergence analysis of discrete time Lion-$\mathcal{K}$. The proof idea is adapted for section B.10, by defining the same Hamiltonian function, we obtain the bound for Δ_t^1 and Δ_t^2.

Compared with the deterministic case, the main challenge is to bound an additional correlation term due to the stochastic gradient at each iteration t:

$$V_t := \mathrm{cov}(g_t,\ \nabla\mathcal{K}(\tilde{m}_{t+1})) = \mathrm{cov}(g_t,\ \nabla\mathcal{K}(\beta_1 m_t + (1-\beta_1) g_t)), \tag{40}$$

where $\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])^\top (Y - \mathbb{E}[Y])]$.

Definition B.11. *For a random variable X on $\mathbb{R}^d$, its (trace of) variance $\mathrm{var}(X)$, when exists, is defined as*

$$\mathrm{var}(X) = \mathbb{E}[\|X - \mathbb{E}[X]\|_2^2]$$

Assumption B.12. *Assume*

$$\mathrm{var}(g_t) \leq \frac{v_{\max}}{n_{batch}},$$

where n_{batch} represents the batch size.

Assumption B.13. *$\mathcal{D}$ is the data distribution, the stochastic sample $\xi_t \sim \mathcal{D}$ is i.i.d., given a function $f(x;\xi)$, the gradient $\nabla f(x;\xi)$ is taken with respect to variable x, and $\mathbb{E}[\nabla f(x,\xi)] = \nabla f(x)$*

Theorem B.14. *Under the assumptions delineated in B.13 and B.12, consider a function $f\colon \mathbb{R}^d \to \mathbb{R}$ that is L-smooth. Additionally, let $\mathcal{K}\colon \mathbb{R}^d \to \mathbb{R}$ be a closed and convex function, consider the following scheme:*

$$\begin{aligned}
m_{t+1} &= \beta_2 m_t - (1-\beta_2) g_t \\
\tilde{m}_{t+1} &= \beta_1 m_t - (1-\beta_1) g_t \\
x_{t+1} &= x_t + \epsilon(\nabla\mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1}),
\end{aligned} \tag{41}$$

where $g_t = \nabla f(x_t;\xi_t)$ *as shown in* B.13, $m_0, g_1, \dots, g_t, \dots$ *are random variables with* $\mathbb{E}[g_t] = \nabla f(x_t)$. $\nabla\mathcal{K}$ *is a weak gradient of* $\mathcal{K}$ *with* $\nabla\mathcal{K}(0) = 0$, $\|\nabla\mathcal{K}(x) - \nabla\mathcal{K}(y)\| \le L_{\mathcal{K}}\|x-y\|, \forall x, y \in \mathbb{R}^d$, *and* $\beta_1, \beta_2 \in (0,1)$, *and* $\beta_2 > \beta_1$, *and* $\epsilon, \lambda > 0$.

Let $\mathcal{K}^*$ *be the conjugate function of* $\mathcal{K}$. *Define the following Lyapunov function:*

$$H(x,m) = f(x) + \frac{1}{\lambda}\mathcal{K}^*(\lambda x) + \frac{\beta_1}{\epsilon\lambda(1-\beta_1) + (1-\beta_2)}(\mathcal{K}^*(\lambda x) + \mathcal{K}(m) - \lambda x^\top m),$$

and

$$\Delta_t^1 = (\nabla\mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top(\tilde{m}_{t+1} - \nabla\mathcal{K}^*(\lambda x_{t+1})),$$
$$\Delta_t^2 = (\nabla\mathcal{K}(\tilde{m}_{t+1}) - \nabla\mathcal{K}(m_{t+1}))^\top(\tilde{m}_{t+1} - m_{t+1}),$$

where $\nabla\mathcal{K}^*$ *is a subgradient of* $\mathcal{K}^*$. *Then we have* $\Delta_t^1 \ge 0$ *and* $\Delta_t^2 \ge 0$ *from Lemma* B.9, *and*

$$\mathbb{E}\left[H(x_{t+1}, m_{t+1}) - H(x_t, m_t)\right] \le \mathbb{E}\left[-\epsilon(a\Delta_t^1 + b\Delta_t^2) + \frac{L\epsilon^2}{2}\left\|\nabla\mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1}\right\|_2^2\right]$$
$$+\epsilon\frac{L_{\mathcal{K}}}{1+\lambda\epsilon}(1-\beta_1)\frac{v_{max}}{n_{batch}} + \frac{L_{\mathcal{K}}}{1+\lambda\epsilon}\sqrt{\frac{(1-\beta_2)}{(1+\beta_2)}}\frac{v_{max}}{n_{batch}}$$

where

$$a = \frac{\epsilon\lambda\beta_1}{\epsilon\lambda(1-\beta_1)+(1-\beta_2)} + 1 \ge 0, \qquad b = \frac{\beta_1(1-\beta_2)}{(\beta_2-\beta_1)(\epsilon\lambda(1-\beta_1)+(1-\beta_2))} \ge 0.$$

$v_{\max}$, n_{batch} *are defined in* B.12

Hence, a telescoping sum yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[a\Delta_t^1 + b\Delta_t^2\right] \le \mathbb{E}\left[\frac{H(x_0,m_0) - H(x_T,m_T)}{\epsilon T} + \frac{L\epsilon}{2}B_t + \frac{C_t}{n_{batch}}\right],$$

where $B_t = \frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1}\|_2^2$, *and* $C_t = \left(\frac{L_{\mathcal{K}}}{1+\lambda\epsilon}(1-\beta_1) + \frac{L_{\mathcal{K}}}{1+\lambda\epsilon}\sqrt{\frac{(1-\beta_2)}{(1+\beta_2)}}\right)v_{max}$.

Proof. The proof is a simple extended variant of 4.1. Following the proof of Theorem B.8, define

$$H(x,m) = (a-c)f(x) + \frac{a}{\lambda}\mathcal{K}^*(\lambda x) + \frac{c}{\lambda}\mathcal{K}(m) - cx^\top m, \quad \text{with} \quad c = (a+b)\beta_1 - b\beta_2,$$

where

$$a = \frac{\epsilon\lambda\beta_1}{\epsilon\lambda(1-\beta_1)+(1-\beta_2)} + 1 \ge 0, \quad b = \frac{\beta_1(1-\beta_2)}{(\beta_2-\beta_1)(\epsilon\lambda(1-\beta_1)+(1-\beta_2))} \ge 0, \quad c = a-1.$$

By the definition of Δ_t^1, Δ_t^2, we have

$$\begin{aligned}
&a\Delta_t^1 + b\Delta_t^2\\
&= a(\nabla\mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top(\tilde{m}_{t+1} - \nabla\mathcal{K}^*(\lambda x_{t+1}))\\
&\qquad + b(\nabla\mathcal{K}(\tilde{m}_{t+1}) - \nabla\mathcal{K}(m_{t+1}))^\top(\tilde{m}_{t+1} - m_{t+1})\\
&= (\nabla\mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top(a(\nabla\mathcal{K}^*(\lambda x_{t+1}) - \tilde{m}_{t+1}) + b(\tilde{m}_{t+1} - m_{t+1}))\\
&\qquad + b(\nabla\mathcal{K}(m_{t+1}) - \lambda x_{t+1})^\top(m_{t+1} - \tilde{m}_{t+1})\\
&= -(\nabla\mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top(a\nabla\mathcal{K}^*(\lambda x_{t+1}) - ((a+b)\beta_1 - b\beta_2)m_t + (a - (a+b)\beta_1 + b\beta_2)\nabla f(x_t))\\
&\qquad - b\frac{\beta_2-\beta_1}{1-\beta_2}\frac{\lambda}{c}(\frac{c}{\lambda}\nabla\mathcal{K}(m_{t+1}) - cx_{t+1})^\top(m_{t+1} - m_t)\\
&= -\left[(a-c)g_t + a\nabla\mathcal{K}^*(\lambda x_{t+1}) - cm_t\right]^\top(\nabla\mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})\\
&\qquad - \frac{1}{\epsilon}\left[\frac{c}{\lambda}\nabla\mathcal{K}(m_{t+1}) - cx_{t+1}\right]^\top(m_{t+1} - m_t)\\
&= -\frac{1}{\epsilon}\left[(a-c)g_t + a\nabla\mathcal{K}^*(\lambda x_{t+1}) - cm_t\right]^\top(x_{t+1} - x_t)\\
&\qquad - \frac{1}{\epsilon}\left[\frac{c}{\lambda}\nabla\mathcal{K}(m_{t+1}) - cx_{t+1}\right]^\top(m_{t+1} - m_t)
\end{aligned} \tag{42}$$

By Lemma B.9,

$$H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \le \hat{\nabla}_x H_t^\top (x_{t+1} - x_t) + \hat{\nabla}_m H_t^\top (m_{t+1} - m_t) + \frac{L}{2} \|x_{t+1} - x_t\|_2^2,$$

where

$$\hat{\nabla}_x H_t = (a-c)\nabla f(x_t) + a\nabla \mathcal{K}^*(\lambda x_{t+1}) - c m_t,$$
$$\hat{\nabla}_m H_t = \frac{c}{\lambda}\nabla \mathcal{K}(m_{t+1}) - c x_{t+1} = \frac{c}{\epsilon\lambda}(\hat{V}_{x,t} - \nabla \mathcal{K}(\tilde{m}_{t+1}) + \nabla \mathcal{K}(m_{t+1}))$$

with

$$V_{x,t} = x_{t+1} - x_t = \epsilon(\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})$$
$$V_{m,t} = m_{t+1} - m_t = -(1-\beta_2)(g_t - m_t)$$
$$\tilde{m}_{t+1} - m_{t+1} = -(\beta_2 - \beta_1)(g_t - m_t) = -(\beta_2 - \beta_1)V_{m,t}$$
$$\hat{V}_{m,t} = -\nabla \mathcal{K}(\tilde{m}_{t+1}) + \nabla \mathcal{K}(m_{t+1})$$

This gives

$$\begin{aligned} &H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \\ &\le \hat{\nabla}_x H_t^\top (x_{t+1} - x_t) + \hat{\nabla}_m H_t^\top (m_{t+1} - m_t) + \frac{L}{2} \|x_{t+1} - x_t\|_2^2 \end{aligned}$$

Hence,

$$\begin{aligned} H(x_{t+1}, m_{t+1}) - H(x_t, m_t) &\le \left[(a-c)\nabla f(x_t) + a\nabla \mathcal{K}^*(\lambda x_{t+1}) - c m_t\right]^\top (x_{t+1} - x_t) \\ &\quad + \left[\frac{c}{\lambda}\nabla \mathcal{K}(m_{t+1}) - c x_{t+1}\right]^\top (m_{t+1} - m_t) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2 \\ &= \left[(a-c)g_t + a\nabla \mathcal{K}^*(\lambda x_{t+1}) - c m_t\right]^\top (x_{t+1} - x_t) \\ &\quad + \left[\frac{c}{\lambda}\nabla \mathcal{K}(m_{t+1}) - c x_{t+1}\right]^\top (m_{t+1} - m_t) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2 \\ &\quad + \epsilon(a-c)(\nabla f(x_t) - g_t)^\top (\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1}) \\ &= -\epsilon(a\Delta_t^1 + b\Delta_t^2) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2 \qquad \text{//by equation 44} \\ &\quad + \epsilon(a-c)(\nabla f(x_t) - g_t)^\top (\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1}) \end{aligned}$$

It suffices to bound $\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top (\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})\right]$.

Note that

$$\begin{aligned} &\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top (\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})\right] \\ &= \mathbb{E}\left[(\nabla f(x_t) - g_t)^\top \left(\frac{1}{1+\lambda\epsilon}\nabla \mathcal{K}(\tilde{m}_{t+1}) - \frac{\lambda}{1+\lambda\epsilon}x_t\right)\right] \\ &= \frac{1}{1+\lambda\epsilon}\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top \nabla \mathcal{K}(\tilde{m}_{t+1})\right] + \frac{\lambda}{1+\lambda\epsilon}\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top x_t\right] \end{aligned}$$

By Assumption B.13,

$$\begin{aligned} \mathbb{E}\left[(\nabla f(x_t) - g_t)^\top \lambda x_t\right] &= \lambda \mathbb{E}_{x_t}\left[\mathbb{E}_{\xi_t}\left[(\nabla f(x_t) - \nabla f(x_t, \xi_t))^\top x_t \mid x_t\right]\right] \\ &= 0 \qquad \text{//by B.13 } \mathbb{E}[\nabla f(x, \xi)] = \nabla f(x) \end{aligned}$$

Next, let us bound $\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top \nabla \mathcal{K}(\tilde{m}_{t+1})\right]$.

$$\begin{aligned} \mathbb{E}\left[(\nabla f(x_t) - g_t)^\top \nabla \mathcal{K}(\tilde{m}_{t+1})\right] &= \mathbb{E}\left[(\nabla f(x_t) - g_t)^\top \nabla \mathcal{K}(\beta_1 m_t - (1-\beta_1)g_t)\right] \\ &\le L_{\mathcal{K}}(1-\beta_1)\mathrm{var}(g_t) + L_{\mathcal{K}}\sqrt{\mathrm{var}(\beta_1 m_t) \cdot \mathrm{var}(g_t)} \qquad \text{//by B.18} \\ &\le L_{\mathcal{K}}(1-\beta_1)\frac{v_{max}}{n_{batch}} + L_{\mathcal{K}}\sqrt{\frac{(1-\beta_2)}{(1+\beta_2)}\frac{v_{max}}{n_{batch}}} \qquad \text{//by B.18} \end{aligned}$$

Hence,

$$
\begin{aligned}
&\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top (\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1})\right] \\
&= \frac{1}{1+\lambda\epsilon}\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top \nabla \mathcal{K}(\tilde{m}_{t+1})\right] + \frac{\lambda}{1+\lambda\epsilon}\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top x_{t+1}\right] \\
&\le \frac{L_{\mathcal{K}}}{1+\lambda\epsilon}(1-\beta_1)\frac{v_{max}}{n_{batch}} + \frac{L_{\mathcal{K}}}{1+\lambda\epsilon}\sqrt{\frac{(1-\beta_2)}{(1+\beta_2)}\frac{v_{max}}{n_{batch}}}
\end{aligned}
$$

□

Lemma B.15. *Let X, Y be two $\mathbb{R}^d$-valued random variables with $\mathrm{var}(X) < +\infty$ and $\mathrm{var}(Y) < +\infty$, and assume $\mathcal{K}$ yields a weak derivative $\nabla\mathcal{K}$. We have*

$$
\mathbb{E}[(Y - \mathbb{E}Y)^\top \nabla\mathcal{K}(X + \epsilon Y)] \le L_{\mathcal{K}}\epsilon \mathrm{var}(Y) + L_{\mathcal{K}}\mathrm{var}(X)\cdot\mathrm{var}(Y)
$$

Proof.

$$
\begin{aligned}
&\mathbb{E}[(Y - \mathbb{E}[Y])^\top \nabla\mathcal{K}(X + \epsilon Y)] \\
&= \mathbb{E}[(Y - \mathbb{E}[Y])^\top \left(\nabla\mathcal{K}(X + \epsilon Y) - \mathcal{K}(\mathbb{E}[X] + \epsilon\mathbb{E}[Y])\right)] \\
&= \sqrt{\mathbb{E}\left\|Y - \mathbb{E}[Y]\right\|^2}\sqrt{\mathbb{E}\left\|\nabla\mathcal{K}(X + \epsilon Y) - \mathcal{K}(\mathbb{E}[X] + \epsilon\mathbb{E}[Y])\right\|^2} \\
&= \sqrt{\mathbb{E}\left\|Y - \mathbb{E}[Y]\right\|^2}\sqrt{L_{\mathcal{K}}\mathbb{E}\left\|X + \epsilon Y - \mathbb{E}[X] - \epsilon\mathbb{E}[Y]\right\|^2} \\
&= L_{\mathcal{K}}\sqrt{\mathbb{E}\left\|Y - \mathbb{E}[Y]\right\|^2}\left(\sqrt{\mathbb{E}\left\|X - \mathbb{E}[X]\right\|^2} + \sqrt{\epsilon^2\left\|Y - \epsilon\mathbb{E}[Y]\right\|^2}\right) \\
&= L_{\mathcal{K}}\epsilon\mathbb{E}\left\|Y - \mathbb{E}[Y]\right\|^2 + L_{\mathcal{K}}\sqrt{\mathbb{E}\left\|Y - \mathbb{E}[Y]\right\|^2}\sqrt{\mathbb{E}\left\|X - \mathbb{E}[X]\right\|^2} \\
&= L_{\mathcal{K}}\epsilon\mathrm{var}(Y) + L_{\mathcal{K}}\sqrt{\mathrm{var}(X)\cdot\mathrm{var}(Y)}
\end{aligned}
$$

□

Lemma B.16 (Cumulative error of stochastic gradient [4]). *Following the same setting in theorem B.14, denote $\delta_l = g_l - \nabla f(x_l)$, for any $k < \infty$ and fixed weight $-\infty < \alpha_1, ..., \alpha_k < \infty$, $\sum_{l=1}^k \delta_l$ is a Martingale. In particular,*

$$
\mathbb{E}\left[\left[\sum_{l=1}^k \alpha_l\delta_l\right]^2\right] \le \sum_{l=1}^k \alpha_l^2\sigma^2.
$$

Proof. We simply check the definition of a Martingale. Denote $Y_k := \sum_{l=1}^k \alpha_l\delta_l$. First, we have that

$$
\begin{aligned}
\mathbb{E}[|Y_k|] &= \mathbb{E}\left[\left|\sum_{l=1}^k \alpha_l\delta_l\right|\right] \\
&\le \sum_l |\alpha_l|\mathbb{E}[|\delta_l|] && \text{triangle inequality} \\
&= \sum_l |\alpha_l|\mathbb{E}[\mathbb{E}[|\delta_l||x_l]] && \text{law of total probability} \\
&\le \sum_l |\alpha_l|\mathbb{E}[\sqrt{\mathbb{E}[\delta_l^2|x_l]}] && \text{Jensen's inequality} \\
&\le \sum_l |\alpha_l|\sigma < \infty
\end{aligned}
$$

Second, again using the law of total probability,

$$\begin{aligned}\mathbb{E}[Y_{k+1}|Y_1,...,Y_k] &= \mathbb{E}\left[\sum_{l=1}^{k+1}\alpha_l\delta_l\middle|\alpha_1\delta_1,...,\alpha_k\delta_k\right]\\ &= Y_k + \alpha_{k+1}\mathbb{E}\left[\delta_{k+1}|\alpha_1\delta_1,...,\alpha_k\delta_k\right]\\ &= Y_k + \alpha_{k+1}\mathbb{E}\left[\mathbb{E}\left[\delta_{k+1}|x_{k+1},\alpha_1\delta_1,...,\alpha_k\delta_k\right]|\alpha_1\delta_1,...,\alpha_k\delta_k\right]\\ &= Y_k + \alpha_{k+1}\mathbb{E}\left[\mathbb{E}\left[\delta_{k+1}|x_{k+1}\right]|\alpha_1\delta_1,...,\alpha_k\delta_k\right]\\ &= Y_k\end{aligned}$$

This completes the proof that it is indeed a Martingale. We now make use of the properties of Martingale difference sequences to establish a variance bound on the Martingale.

$$\begin{aligned}\mathbb{E}[[\sum_{l=1}^{k}\alpha_l\delta_l]^2] &= \sum_{l=1}^{k}\mathbb{E}[\alpha_l^2\delta_l^2] + 2\sum_{l<j}\mathbb{E}[\alpha_l\alpha_j\delta_l\delta_j]\\ &= \sum_{l=1}^{k}\alpha_l^2\mathbb{E}[\mathbb{E}[\delta_l^2|\delta_1,...,\delta_{l-1}]] + 2\sum_{l<j}\alpha_l\alpha_j\mathbb{E}\Big[\delta_l\mathbb{E}\big[\mathbb{E}[\delta_j|\delta_1,...,\delta_{j-1}]\big|\delta_l\big]\Big]\\ &= \sum_{l=1}^{k}\alpha_l^2\mathbb{E}[\mathbb{E}[\mathbb{E}[\delta_l^2|x_l,\delta_1,...,\delta_{l-1}]|\delta_1,...,\delta_{l-1}]] + 0\\ &= \sum_{l=1}^{k}\alpha_l^2\sigma^2.\end{aligned}$$

□

The consequence of this lemma is that we are able to treat $\delta_1,...,\delta_k$ as if they are independent, even though they are not—clearly δ_l is dependent on $\delta_1,...,\delta_{l-1}$ through x_l. By Lemma B.16, we can compute the variance of momentum m_t,

$$\begin{aligned}\mathrm{var}(m_t) &= (1-\beta_2)^2\mathbb{E}\left\|\sum_{i=1}^{t}\beta_2^{t-i}\delta_i\right\|^2\\ &= (1-\beta_2)^2\mathbb{E}\sum_{i=1}^{t}\beta_2^{2t-2i}\left\|\delta_i\right\|^2\\ &= \frac{(1-\beta_2)v_{max}}{(1+\beta_2)n_{batch}}\end{aligned}$$

B.12 Analysis with Stochastic Gradient LION

Theorem B.17. *Under the assumptions delineated in B.13 and B.12, consider a function $f:\mathbb{R}^d\to\mathbb{R}$ that is L-smooth. Consider the following scheme:*

$$\begin{aligned}m_{t+1} &= \beta_2 m_t - (1-\beta_2)g_t\\ \tilde{m}_{t+1} &= \beta_1 m_t - (1-\beta_1)g_t\\ x_{t+1} &= x_t + \epsilon(sign(\tilde{m}_{t+1}) - \lambda x_{t+1}),\end{aligned} \tag{43}$$

where $g_t = \nabla f(x_t;\xi_t)$ as shown in B.13, $m_0, g_1, \ldots, g_t, \ldots$ are random variables with $\mathbb{E}[g_t] = \nabla f(x_t)$. $\beta_1, \beta_2 \in (0,1)$, and $\beta_2 > \beta_1$, and $\epsilon, \lambda > 0$.

Define the following Lyapunov function:

$$H(x,m) = f(x) + \frac{1}{\lambda}\|\lambda x\|^* + \frac{\beta_1}{\epsilon\lambda(1-\beta_1)+(1-\beta_2)}(\|\lambda x\|^* + \|m\| - \lambda x^\top m),$$

and

$$\begin{aligned}\Delta_t^1 &= (sign(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top(\tilde{m}_{t+1} - sign^*(\lambda x_{t+1})),\\ \Delta_t^2 &= (sign(\tilde{m}_{t+1}) - sign(m_{t+1}))^\top(\tilde{m}_{t+1} - m_{t+1}),\end{aligned}$$

where $sign^*$ is a subgradient of $\mathcal{K}^*$. Then we have $\Delta_t^1 \geq 0$ and $\Delta_t^2 \geq 0$ from Lemma B.9, and

$$\mathbb{E}\left[H(x_{t+1}, m_{t+1}) - H(x_t, m_t)\right] \leq \mathbb{E}\left[-\epsilon(a\Delta_t^1 + b\Delta_t^2) + \frac{L\epsilon^2}{2}\left\|sign(\tilde{m}_{t+1}) - \lambda x_{t+1}\right\|_2^2\right] + \epsilon\frac{1}{1+\lambda\epsilon}\frac{\sqrt{d \cdot v_{max}}}{\sqrt{n_{batch}}}$$

where

$$a = \frac{\epsilon\lambda\beta_1}{\epsilon\lambda(1-\beta_1) + (1-\beta_2)} + 1 \geq 0, \qquad b = \frac{\beta_1(1-\beta_2)}{(\beta_2-\beta_1)(\epsilon\lambda(1-\beta_1)+(1-\beta_2))} \geq 0.$$

v_{max}, n_{batch} are defined in B.12

Hence, a telescoping sum yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[a\Delta_t^1 + b\Delta_t^2\right] \leq \mathbb{E}\left[\frac{H(x_0, m_0) - H(x_T, m_T)}{\epsilon T} + \frac{L\epsilon}{2}B_t + \frac{1}{1+\lambda\epsilon}\frac{\sqrt{d \cdot v_{max}}}{\sqrt{n_{batch}}}\right],$$

where $B_t = \frac{1}{T}\sum_{t=1}^{T}\left\|sign(\tilde{m}_{t+1}) - \lambda x_{t+1}\right\|_2^2$

Proof. Define

$$H(x, m) = (a-c)f(x) + \frac{a}{\lambda}\|\lambda x\|^* + \frac{c}{\lambda}\|m\| - cx^\top m, \qquad \text{with} \qquad c = (a+b)\beta_1 - b\beta_2,$$

where

$$a = \frac{\epsilon\lambda\beta_1}{\epsilon\lambda(1-\beta_1) + (1-\beta_2)} + 1 \geq 0, \quad b = \frac{\beta_1(1-\beta_2)}{(\beta_2-\beta_1)(\epsilon\lambda(1-\beta_1)+(1-\beta_2))} \geq 0, \quad c = a - 1.$$

By the definition of Δ_t^1, Δ_t^2, we have

$$\begin{aligned}
&a\Delta_t^1 + b\Delta_t^2 \\
&= a(sign(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top(\tilde{m}_{t+1} - sign^*(\lambda x_{t+1})) \\
&\qquad + b(sign(\tilde{m}_{t+1}) - sign(m_{t+1}))^\top(\tilde{m}_{t+1} - m_{t+1}) \\
&= (sign(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top(a(sign^*(\lambda x_{t+1}) - \tilde{m}_{t+1}) + b(\tilde{m}_{t+1} - m_{t+1})) \\
&\qquad + b(sign(m_{t+1}) - \lambda x_{t+1})^\top(m_{t+1} - \tilde{m}_{t+1}) \\
&= -(sign(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top(asign^*(\lambda x_{t+1}) - ((a+b)\beta_1 - b\beta_2)m_t + (a - (a+b)\beta_1 + b\beta_2)\nabla f(x_t)) \\
&\qquad - b\frac{\beta_2-\beta_1}{1-\beta_2}\frac{\lambda}{c}(\frac{c}{\lambda}sign(m_{t+1}) - cx_{t+1})^\top(m_{t+1} - m_t) \\
&= -\left[(a-c)g_t + asign^*(\lambda x_{t+1}) - cm_t\right]^\top(sign(\tilde{m}_{t+1}) - \lambda x_{t+1}) \\
&\qquad - \frac{1}{\epsilon}\left[\frac{c}{\lambda}sign(m_{t+1}) - cx_{t+1}\right]^\top(m_{t+1} - m_t) \\
&= -\frac{1}{\epsilon}\left[(a-c)g_t + asign^*(\lambda x_{t+1}) - cm_t\right]^\top(x_{t+1} - x_t) \\
&\qquad - \frac{1}{\epsilon}\left[\frac{c}{\lambda}sign(m_{t+1}) - cx_{t+1}\right]^\top(m_{t+1} - m_t)
\end{aligned} \tag{44}$$

By Lemma B.9,

$$H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \leq \hat{\nabla}_x H_t^\top(x_{t+1} - x_t) + \hat{\nabla}_m H_t^\top(m_{t+1} - m_t) + \frac{L}{2}\left\|x_{t+1} - x_t\right\|_2^2,$$

where

$$\begin{aligned}
\hat{\nabla}_x H_t &= (a-c)\nabla f(x_t) + asign^*(\lambda x_{t+1}) - cm_t, \\
\hat{\nabla}_m H_t &= \frac{c}{\lambda}sign(m_{t+1}) - cx_{t+1} = \frac{c}{\epsilon\lambda}(\hat{V}_{x,t} - sign(\tilde{m}_{t+1}) + sign(m_{t+1}))
\end{aligned}$$

with

$$\begin{aligned}
&V_{x,t} = x_{t+1} - x_t = \epsilon(sign(\tilde{m}_{t+1}) - \lambda x_{t+1}) \\
&V_{m,t} = m_{t+1} - m_t = -(1-\beta_2)(g_t - m_t) \\
&\tilde{m}_{t+1} - m_{t+1} = -(\beta_2 - \beta_1)(g_t - m_t) = -(\beta_2 - \beta_1)V_{m,t} \\
&\hat{V}_{m,t} = -sign(\tilde{m}_{t+1}) + sign(m_{t+1})
\end{aligned}$$

This gives

$$\begin{aligned}
&H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \\
&\leq \hat{\nabla}_x H_t^\top (x_{t+1} - x_t) + \hat{\nabla}_m H_t^\top (m_{t+1} - m_t) + \frac{L}{2} \|x_{t+1} - x_t\|_2^2
\end{aligned}$$

Hence,

$$\begin{aligned}
H(x_{t+1}, m_{t+1}) - H(x_t, m_t) &\leq \left[(a-c)\nabla f(x_t) + asign^*(\lambda x_{t+1}) - cm_t\right]^\top (x_{t+1} - x_t) \\
&\quad + \left[\frac{c}{\lambda} sign(m_{t+1}) - cx_{t+1}\right]^\top (m_{t+1} - m_t) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2 \\
&= \left[(a-c)g_t + asign^*(\lambda x_{t+1}) - cm_t\right]^\top (x_{t+1} - x_t) \\
&\quad + \left[\frac{c}{\lambda} sign(m_{t+1}) - cx_{t+1}\right]^\top (m_{t+1} - m_t) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2 \\
&\quad + \epsilon(a-c)(\nabla f(x_t) - g_t)^\top (sign(\tilde{m}_{t+1}) - \lambda x_{t+1}) \\
&= -\epsilon(a\Delta_t^1 + b\Delta_t^2) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2 \qquad \text{//by equation 44} \\
&\quad + \epsilon(a-c)(\nabla f(x_t) - g_t)^\top (sign(\tilde{m}_{t+1}) - \lambda x_{t+1})
\end{aligned}$$

It suffices to bound $\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top (sign(\tilde{m}_{t+1}) - \lambda x_{t+1})\right]$.

Note that

$$\begin{aligned}
&\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top (sign(\tilde{m}_{t+1}) - \lambda x_{t+1})\right] \\
&= \mathbb{E}\left[(\nabla f(x_t) - g_t)^\top (\frac{1}{1+\lambda\epsilon} sign(\tilde{m}_{t+1}) - \frac{\lambda}{1+\lambda\epsilon} x_t)\right] \\
&= \frac{1}{1+\lambda\epsilon}\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top sign(\tilde{m}_{t+1})\right] + \frac{\lambda}{1+\lambda\epsilon}\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top x_t\right]
\end{aligned}$$

By Assumption B.13,

$$\begin{aligned}
\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top \lambda x_t\right] &= \lambda \mathbb{E}_{x_t}\left[\mathbb{E}_{\xi_t}\left[(\nabla f(x_t) - \nabla f(x_t, \xi_t))^\top x_t \mid x_t\right]\right] \\
&= 0 \qquad \text{//by B.13 } \mathbb{E}[\nabla f(x, \xi)] = \nabla f(x)
\end{aligned}$$

Next, we can use B.18 to bound $\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top sign(\tilde{m}_{t+1})\right]$.

$$\begin{aligned}
\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top sign(\tilde{m}_{t+1})\right] &= \mathbb{E}\left[(\nabla f(x_t) - g_t)^\top sign(\beta_1 m_t - (1-\beta_1)g_t)\right] \\
&\leq \sqrt{d \cdot \mathrm{var}(g_t)} \qquad \text{//by B.18} \\
&\leq \sqrt{\frac{d \cdot v_{max}}{n_{batch}}} \qquad \text{//by B.12}
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top (sign(\tilde{m}_{t+1}) - \lambda x_{t+1})\right] \\
&= \frac{1}{1+\lambda\epsilon}\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top sign(\tilde{m}_{t+1})\right] + \frac{\lambda}{1+\lambda\epsilon}\mathbb{E}\left[(\nabla f(x_t) - g_t)^\top x_{t+1}\right] \\
&\leq \frac{1}{1+\lambda\epsilon}\sqrt{\frac{d \cdot v_{max}}{n_{batch}}}
\end{aligned}$$

□

Lemma B.18. *Let X, Y be two $\mathbb{R}^d$-valued random variables with $\operatorname{var}(Y) < +\infty$, and assume $\mathcal{K}$ yields a weak derivative sign. We have $\mathbb{E}[(Y - \mathbb{E}Y)^\top sign(X + \epsilon Y)] \leq \sqrt{d \operatorname{var}(Y)}$*

Proof.

$$\mathbb{E}[(Y - \mathbb{E}[Y])^\top sign(X + \epsilon Y)] \leq \mathbb{E}[|Y - \mathbb{E}[Y]|] \leq \sqrt{d \cdot \mathbb{E}[\|Y - \mathbb{E}[Y]\|^2]} = \sqrt{d \cdot \operatorname{var}(Y)}$$

□