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 \( \{ z = 0 \} \):

\[
\delta(z) = \begin{cases} 
0 & \text{if } z = 0 \\
+\infty & \text{if } z \neq 0.
\end{cases}
\]

\( \mathbb{I}(z) = \begin{cases} 
0 & \text{if } z = 0 \\
1 & \text{if } z \neq 0.
\end{cases} \)

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

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

\[
\nabla \mathcal{K}(x) = \frac{\text{sign}(x) |x|^{p-1}}{\|x\|_p^{p-1}},
\]

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 \( \ell_p \) norm correspond to solving

\[
\min_x f(x) \quad \text{s.t.} \quad \|x\|_q \leq 1/\lambda.
\]

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

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

The conjugate function is

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

Hence, Lion-\( \mathcal{K} \) with grouped \( \ell_p \) norm corresponds to solving

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

Lower Truncated \( \ell_1 \) Norm Consider \( \mathcal{K}(x) = \sum_{i=1}^d \max(|x_i| - c, 0) \) where \( c > 0 \). We can take

\[
\nabla \mathcal{K}(x) = \mathbb{I}(|x| \geq e) \text{sign}(x), \quad (16)
\]

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

\[
\mathcal{K}^*(x) = \sup_z \sum_{i=1}^d (x_i z_i - \max(|z_i| - e, 0))
\]

\[
= \sup_z, c \sum_{i=1}^d (x_i z_i - c_i) \quad \text{s.t.} \quad c_i \geq 0, \quad c \geq |z_i| - e
\]

\[
= \sup_{c \geq 0} \sum_{i=1}^d |x_i| (c_i + e) - c_i
\]

\[
= \sum_{i=1}^d \delta(|x_i| \leq 1) + e |x_i|
\]

\[
= \delta(\|x\|_\infty \leq 1) + e \|x\|_1.
\]

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

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

Hence, truncating the small gradients in Lion induces an \( \ell_1 \) penalty, which encourages the sparsity of the final solution.
**Lower (Vector-wise) Truncated $\ell_p$ Norm**  Consider $\mathcal{K}(x) = \max(\|x\|_p - e, 0)$. We have

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

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

$$\mathcal{K}^*(x) = \sup_{x^\top z = \max(\|z\|_p - e, 0)} (z^\top x - c) s.t. c \geq 0, c \geq \|z\|_p - e$$

$$= \sup_{c \geq 0} \|x\|_q (c + e) - c$$

$$= \delta(\|x\|_q \leq 1) + e \|x\|_q.$$

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

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

**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

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

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

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

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

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

where $\text{rank}(i, x)$ denotes the rank of $|x_i|$ in $x$.

$$\mathcal{K}^*(x) = \sup_z \left\{ x^\top z - \sum_i c_i |z(i)| \right\}$$

$$= \sup_{z \geq 0} \left\{ \sum_i |x(i)| \times z(i) - \sum_i c_i z(i) \right\} \text{ by rearrangement inequality}$$

$$= \sup_{w \geq 0} \left\{ \sum_i (|x(i)| - c_i) \times (\sum_j w_j) \right\} \text{ let } z(i) = \sum_{j \geq i} w_j, w_j \geq 0$$

$$= \sup_{w \geq 0} \left\{ \sum_j \sum_{i \leq j} (|x(i)| - c_i) \times w_j \right\} \text{ let } z(i) = \sum_{j \geq i} w_j, w_j \geq 0$$

$$= \sum_j \delta(\sum_{i \leq j} |x(i)| \leq \sum_{j \leq i} c_j)$$

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

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

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

$$\mathcal{K}(x) = \sum_{i \leq i^{\text{cut}}} |x(i)|, \quad \nabla \mathcal{K}(x) = I(|x| \geq x_{i^{\text{cut}}}) \text{sign}(x),$$
in which we zero out the updates of the elements whose absolute values are smaller than the $i^{\text{cut}}$-th largest element. It is useful to compare this with (16) which applies the truncation based on a fixed number $\epsilon$, rather than the percentile.

The conjugate is

$$K^*(x) = \sum_{j \leq i^{\text{cut}}} \delta(|x(j)| \leq 1) + \delta(\|x\|_1 \leq i^{\text{cut}})$$

Then, Lion-$K$ in this case corresponds to solving

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

in which the percentile-based truncation effectively imposes a constraint on the $\ell_1$ norm of $x$. It is different from (17) in which the $\ell_1$ norm appears as a regularization term in the objective, rather than as a hard constraint.

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

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

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

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

Lion-$K$ correspond to solving an entropy-regularized optimization:

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

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

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

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

We have

$$\nabla K(x) = \text{Clip}(x, -a, a) / a,$$

with

$$\text{Clip}(x, a, b) = \begin{cases} x_i & \text{if } x_i \in [a, b] \\ b & \text{if } x > b \\ a & \text{if } x < a. \end{cases}$$

The conjugate is

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

$$K^*(x) = \sum_{i=1}^d \max\left( \sup_{|z_i| \geq a} x_i z_i - |z_i|, \sup_{|z_i| < a} x_i z_i - \frac{1}{2a} z_i^2 \right)$$

$$= \sum_{i=1}^d \max\left( \delta(|x_i| \leq 1) + a(|x_i| - 1), \frac{1}{2} a x_i^2 \right)$$

$$= \sum_{i=1}^d \delta(|x_i| \leq 1) + \frac{1}{2} a x_i^2$$

$$= \frac{a}{2} \|x\|_2^2 + \delta(\|x\|_\infty \leq 1).$$
Relativistic Consider \( K(x) = \sum_{i=1}^{d} \sqrt{x_i^2 + e^2} \), then \( \nabla K(x) = \frac{x}{\sqrt{x^2 + e^2}} \), and

\[
K^*(x) = \sup_{z} \left( \sum_{i=1}^{d} x_i z_i - \sqrt{z_i^2 + e^2} \right)
\]

\[
\begin{align*}
&= \sum_{i=1}^{d} \frac{x_i^2 e}{1 - x_i^2} - \frac{e}{\sqrt{1 - x_i^2}} \\
&= \sum_{i=1}^{d} -e \sqrt{1 - x_i^2} + \delta(|x_i| \leq 1) \\
&= \sum_{i=1}^{d} -e \sqrt{1 - x_i^2} + \delta(||x||_{\infty} \leq 1)
\end{align*}
\]

A related case is

\[
K(x) = |x| - e \log(|x|/e + 1), \quad \text{with} \quad \nabla K(x) = \frac{x}{|x| + e},
\]

whose conjugate function is

\[
K^*(x) = \sup_{x} \left( \sum_{i=1}^{d} x_i z_i - |z_i| + e \log(|z_i|/e + 1) \right)
\]

\[
\begin{align*}
&= \sum_{i=1}^{d} |x_i|^2 e/(1 - |x_i|) - |x_i| e/(1 - |x_i|) + e \log(1/(1 - |x_i|)) \\
&= \sum_{i=1}^{d} -e(|x_i| + \log(1 - |x_i|)) + \delta(||x||_{\infty} < 1)
\end{align*}
\]

B Proofs

B.1 Convex Function Preliminaries

Lemma 2.1 Assume \( K, K^* \) is a closed convex conjugate pair and \( \nabla K, \nabla K^* \) are their subgradients, we have

\[
(\nabla K(x) - \nabla K(y))^T (x - y) \geq 0, \quad (\nabla K(x) - y)^T (x - \nabla K^*(y)) \geq 0. \tag{18}
\]

Proof. 1) By definition of subgradient, we have

\[
K(y) - K(x) \geq \nabla K(x)^T (y - x)
\]

\[
K(x) - K(y) \geq \nabla K(y)^T (x - y).
\]

Summing them together yields \((\nabla K(x) - \nabla K(y))^T (x - y) \geq 0\).

2) Because \( \nabla K^*(y) \in \partial K^*(y) \), we have

\[
K^*(\nabla K(x)) - K^*(y) \geq \nabla K^*(y)^T (\nabla K(x) - y),
\]

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

\[
K^*(y) - K^*(\nabla K(x)) \geq x^T (y - \nabla K(x)).
\]

Summing the two inequalities above yields

\[
(\nabla K(x) - y)^T (\nabla K^*(y) - y) \leq (K^*(\nabla K(x)) - K^*(y)) + (K^*(y) - K^*(\nabla K(x))) = 0.
\]

\[\square\]
B.2 Connection with Nesterov Momentum

Lemma B.1. The Lion-\(K\) ODE is

\[
\begin{align*}
\dot{x}_t &= \nabla K(m_t - \varepsilon(\alpha \nabla f(x_t) + \gamma m_t)) - \lambda x_t \\
\dot{m}_t &= -\alpha \nabla f(x_t) - \gamma m_t.
\end{align*}
\]

is equivalent to

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

if \(K^*\) and \(f\) are second order differentiable.

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

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

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 \(\lambda\) and momentum damping coefficient \(\gamma\) is symmetric in (21).

2) When either the weight decay or momentum damping is turned off, i.e., \(\gamma \lambda = 0\), the \(\ell_2\) regularization in \(F(x)\) is turned off, and we have

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

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

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

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

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

Plugging it into the second equation yields

\[
(1 - \varepsilon \gamma) \dot{m}_t = -\alpha (1 - \varepsilon \gamma) \nabla f(x_t) - \gamma (\nabla K^*(\dot{x}_t + \lambda x_t) + \varepsilon \alpha \nabla f(x_t))
\]

\[
= -\alpha \nabla f(x_t) - \gamma \nabla K^*(\dot{x}_t + \lambda x_t). \tag{23}
\]

Combining (22) and (23) yields

\[
\frac{d}{dt} (\nabla K^*(\dot{x}_t + \lambda x_t) + \varepsilon \alpha \nabla f(x_t)) = -\alpha \nabla f(x_t) - \gamma \nabla K^*(\dot{x}_t + \lambda x_t).
\]

Or

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

\(\square\)

B.3 Discrete-time Schemes of Lion-\(K\)

In the most general form, the Euler approximation of the Lion-\(K\) ODE with step size \(\epsilon\) is

\[
\begin{align*}
x_{t+1} &= x_t + \epsilon(\nabla K(m_t - \varepsilon(\alpha \nabla f(x_t) + \gamma m_t)) - \lambda x_t) \\
m_{t+1} &= m_t - \epsilon(\alpha \nabla f(x_t) + \gamma m_t), \tag{24}
\end{align*}
\]
The discrete Lion-K scheme in (2) is recovered when \( \alpha = \gamma, \beta_1 = 1 - \varepsilon \gamma, \beta_2 = 1 - \varepsilon \gamma. \) 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-K:

\[
\text{Undamped Lion-K: }\ x_{t+1} = x_t + \varepsilon \nabla K(m_t - \beta_1 \nabla f(x_t)) - \lambda x_t \\
m_{t+1} = m_t - \beta_2 \nabla f(x_t),
\]

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

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

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

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

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

Hence,

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

When \( \nabla K^*(x) = x, \) we have

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

It is simplified into

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

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

\[
\dot{x}_t = \nabla K(-\nabla f(x_t)) - \lambda x_t,
\]

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

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

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

\[
\frac{d}{dt} F(x_t) = (\nabla f(x) + \nabla K^*(\lambda x))^\top (\nabla K(-\nabla f(x)) - \lambda x) \leq 0,
\]

where the inequality follows Lemma 2.1.

The Euler discretization of (25) is

\[
x_{t+1} = x_t + \varepsilon (\nabla K(-\nabla f(x_t)) - \lambda x_t).
\]

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

\[
y_{t+1} = \arg \min_x \left\{ \nabla f(x_t)^\top (x - x_t) + \frac{1}{\lambda} K^*(\lambda x) \right\}
\]

\[
x_{t+1} = x_t + \varepsilon_0 (y_{t+1} - x_t),
\]

Solving \( y_{t+1} \) yields

\[
y_{t+1} = \frac{1}{\lambda} \nabla K(-\nabla f(x_t)), \quad \text{and hence} \quad x_{t+1} = (1 - \varepsilon_0) x_t + \frac{\varepsilon_0}{\lambda} \nabla K(-\nabla f(x_t)).
\]

Taking \( \varepsilon = \varepsilon_0 \alpha \) yields (26).
Dual Space Preconditioning and Mirror Descent

When we further set $\lambda = 0$ in (26), Lion-$\mathcal{K}$ reduces to

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

(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

$$x_{t+1} = x_t + e\delta_t,$$

$$\delta_t = \arg\min_\delta \left\{ \nabla f(x_t)^\top \delta + \mathcal{K}^*(\delta) \right\}.$$

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 $B_h(x \mid y)$ is the Bregman divergence associated with a convex function $h: \mathbb{R}^d \to \mathbb{R}$ is defined as

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

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

$$\mathcal{K}^*(\delta) - \mathcal{K}^*(0) = B_{\mathcal{K}^*}(\delta \mid 0) = B_{\mathcal{K}^*}(x_t + e\delta \mid x_t),$$

where $\mathcal{K}^* = \mathcal{K}^*(\frac{\delta - x_t}{\epsilon})$.

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

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

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

(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

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

$$\nabla_m H(x, m) = \nabla \mathcal{K}(m) - \lambda x,$$

and (28) can be written into

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

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

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

Then

$$\frac{d}{dt} H(x_t, m_t) = \nabla_x H^\top V_x + \nabla_m H^\top V_m$$

$$= \nabla_x H^\top (\nabla_m H) + \nabla_m H^\top (-\nabla_x H - \dot{H}_m) = -\nabla_m H^\top \dot{H}_m \leq 0.$$

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

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

where the Hamiltonian component is orthogonal to the gradient $[\nabla_x H, \nabla_m H]$ of $H(x, m)$ and preserves the total energy $\dot{H}(x, m)$, and the descent component introduces a damping like effect to decrease the energy $\dot{H}(x, m)$. \qed
B.6 Lion-K without Weight Decay – A Hamiltonian + Descent Derivation

When the weight decay in Lion-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-K ODE is of the following form when $\lambda = 0$:

$$
\begin{align*}
\dot{x}_t &= \nabla K(\tilde{m}_t), \\
\dot{m}_t &= -\alpha \nabla f(x_t) + \gamma m_t
\end{align*}
$$

(29)

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

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

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

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

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

$$
\nabla_x H(x_t, m_t) \tilde{\nabla}_x H_t = \alpha \nabla f(x) \tilde{\nabla}_x H_t
= -\frac{1}{\varepsilon} (\tilde{m}_t^0 - \tilde{m}_t) \nabla K(\tilde{m}_t^0) - \nabla K(\tilde{m}_t) \leq 0,
$$

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

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

$$
\nabla_m H(x_t, m_t) \tilde{\nabla}_m H_t = -\gamma \nabla K((1-\varepsilon \gamma)m_t) \nabla K(0) \tilde{m}_t = \frac{\gamma}{1-\varepsilon \gamma} (K(0) - K((1-\varepsilon \gamma)m_t)) \leq 0.
$$

Hence, we have

$$
\frac{d}{dt} H(x_t, m_t) = \nabla_x H(x_t, m_t) \tilde{\nabla}_x H_t + \nabla_m H(x_t, m_t) \tilde{\nabla}_m H_t
= -\frac{1}{\varepsilon} (\tilde{m}_t^0 - \tilde{m}_t) \nabla K(\tilde{m}_t^0) - \nabla K(\tilde{m}_t) - \gamma \nabla K((1-\varepsilon \gamma)m_t) \nabla K(0) \tilde{m}_t \leq 0.
$$

Moreover, if $m = 0$ is the unique minimum of $K$, and $\varepsilon \gamma < 1$, then $\nabla K((1-\varepsilon \gamma)m_t) \nabla K(0) \tilde{m}_t = 0$ 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-K ODE

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

$$
\begin{align*}
\dot{x}_t &= \nabla K(\tilde{m}_t) - \lambda x_t, \\
\dot{m}_t &= m_t - \varepsilon (\gamma m_t + \alpha \nabla f(x_t)),
\end{align*}
$$

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

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

...
Then $H$ yields a Lyapunov function in that
\[ -\frac{d}{dt} H(x_t, m_t) = \Delta(x_t, m_t) := \frac{\lambda + \gamma}{1 + \varepsilon\lambda} \Delta_1(x_t, m_t) + \frac{1 - \varepsilon\gamma}{(1 + \varepsilon\lambda)} \Delta_2(m_t, \tilde{m}_t) \geq 0, \]
where
\[ \Delta_1(x, \tilde{m}) = (\tilde{m} - \nabla K^*(\lambda x))^T (\nabla K(\tilde{m}) - \lambda x), \]
\[ \Delta_2(m, \tilde{m}) = \frac{1}{\varepsilon} (\tilde{m} - m)^T (\nabla K(\tilde{m}) - \nabla K(m)). \]
Moreover, the accumulation points of all trajectories are stationary points of $F(x) = \alpha f(x) + \frac{\lambda}{2} K^*(\lambda x)$.

**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 examining what inequalities we can write down using the monotonicity of $\nabla K$ and $\nabla K^*$ via Lemma 2.1, and then work out the Lyapunov function backward.

Write $\tilde{m} = m - \varepsilon(\gamma m + \alpha \nabla f(x))$. Because $\nabla K$ is a monotonic mapping, we have by Lemma 2.1 the following key inequalities:
\[ (-\tilde{m} + \nabla K^*(\lambda x))^T (\nabla K(\tilde{m}) - \lambda x) \leq 0, \]
\[ (m - \tilde{m})^T (\nabla K(\tilde{m}) - \nabla K(m)) \leq 0, \]
or equivalently
\[ (\varepsilon \alpha \nabla f(x) - (1 - \varepsilon\gamma)m + \nabla K^*(\lambda x))^T (\nabla K(\tilde{m}) - \lambda x) \leq 0 \tag{31} \]
\[ \varepsilon (\alpha \nabla f(x) + \gamma m)^T (\nabla K(\tilde{m}) - \lambda x) - (\nabla K(m) - \lambda x) \leq 0 \tag{32} \]
Write $V_x = \nabla K(\tilde{m}) - \lambda x$, 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
\[ (\varepsilon \alpha \nabla f(x) - (1 - \varepsilon\gamma)m + \nabla K^*(\lambda x))^T V_x \leq 0 \tag{33} \]
\[ -\varepsilon V_m^T (V_x - (\nabla K(m) - \lambda x)) \leq 0 \tag{34} \]
Taking $\frac{1}{\varepsilon(1 + \eta)} (Eq. (33) + \eta \times Eq. (34))$ for any $\eta \geq 0$, we get
\[ \left( \alpha \nabla f(x) - \frac{1 - \varepsilon\gamma(1 + \eta)}{\varepsilon(1 + \eta)} m + \frac{1}{\varepsilon(1 + \eta)} \nabla K^*(\lambda x) \right)^T V_x + \frac{\eta \varepsilon}{\varepsilon(1 + \eta)} (\nabla K(m) - \lambda x)^T V_m \leq 0 \]
Define
\[ \tilde{H}(x, m) = \alpha f(x) + \frac{1}{\varepsilon(1 + \eta)} \nabla K^*(\lambda x) + \frac{1 - \varepsilon\gamma(1 + \eta)}{\varepsilon(1 + \eta)} \frac{1}{\lambda} K(m) - \frac{1 - \varepsilon\gamma(1 + \eta)}{\varepsilon(1 + \eta)} m^T x. \]
Then the inequality was reduced to
\[ \nabla_x \tilde{H}(x, m)^T V_x + \frac{\varepsilon \eta \lambda}{1 - \varepsilon\gamma(1 + \eta)} \nabla_m \tilde{H}(x, m)^T V_m \leq 0. \]
If we take $\eta$ such that
\[ \frac{\varepsilon \eta \lambda}{1 - \varepsilon\gamma(1 + \eta)} = 1, \tag{35} \]
then we have when following $\dot{x} = V_x$ and $\dot{m} = V_m$,
\[ \frac{d}{dt} \tilde{H}(x, m) = \nabla_x \tilde{H}(x, m)^T V_x + \nabla_m \tilde{H}(x, m)^T V_m \leq 0. \]
Furthermore, when (35) holds, we have
\[ \eta = \frac{1 - \varepsilon\gamma}{\varepsilon(\lambda + \gamma)}, \quad \frac{1}{\varepsilon(1 + \eta)} = \frac{\lambda + \gamma}{1 + \varepsilon\lambda}, \quad \frac{1 - \varepsilon\gamma(1 + \eta)}{\varepsilon(1 + \eta)} = \frac{(1 - \varepsilon\gamma)\lambda}{1 + \varepsilon\lambda}, \tag{36} \]
Lemma B.4. We provide the decomposition structure $(f + H(x, m))$. Hence, all trajectories in $I$ are also constants in the trajectories in $I$. This suggests that $x_t$ is constant for the trajectories in $I$. Because $\dot{m}_t = \nabla K^*(\lambda x_t)$ and $\dot{m}_t = m_t - \varepsilon(\alpha \nabla f(x_t) + \gamma m_t)$, we have

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

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

$$0 = (1 - \varepsilon \gamma) \dot{m}_t$$

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

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

$$\nabla x H(x, m) = -\dot{V}_x(x, m) + \eta V_m(x, m)$$

$$\nabla m H(x, m) = -\dot{V}_m(x, m) - \eta V_x(x, m),$$

and hence

$$\dot{H}(x, m) = \alpha f(x) + \frac{\lambda + \gamma}{(1 + \varepsilon \lambda)} - \frac{1 - \varepsilon \gamma}{1 + \varepsilon \lambda} \nabla K^*(\lambda x) + \frac{1 - \varepsilon \gamma}{1 + \varepsilon \lambda} (K^*(\lambda x) + K(m) - \lambda m^T x)$$

$$= \alpha f(x) + \frac{\gamma}{\lambda} K^*(\lambda x) + \frac{1 - \varepsilon \gamma}{1 + \varepsilon \lambda} (K^*(\lambda x) + K(m) - \lambda m^T x)$$

$$= H(x, m).$$

In this case,

$$\frac{d}{dt} H(x, m)$$

$$= \frac{1}{\varepsilon(1 + \eta)} (Eq. (33) + \eta \times Eq. (34))$$

$$= \frac{\lambda + \gamma}{1 + \varepsilon \lambda} \nabla K^*(\lambda x)\nabla K(m)$$

$$= \frac{1 - \varepsilon \gamma}{1 + \varepsilon \lambda} \nabla K^*(\lambda x)$$

$$= 0.$$

To ensure that $\eta \geq 0$, we need $\varepsilon \gamma \leq 1.$

LaSalle’s invariance principle Let $H(z)$ is a continuously differentiable Lyapunov function of $\frac{d}{dt} z_t = v(z_t)$, satisfying $\frac{d}{dt} H(z_t) \leq 0$. By LaSalle’s Invariance Principle, the accumulation points of any trajectories of $\frac{d}{dt} z_t = v(z_t)$ is included in

$$I = \{ \text{the union of all trajectories } z_t, \text{satisfying } \frac{d}{dt} H(z_t) = 0 \text{ for all } t \geq 0 \}.$$
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,
\[
\hat{V}_x(x, m)^\top V_x(x, m) \geq 0, \quad \hat{V}_m(x, m)^\top V_m(x, m) \geq 0, \quad \forall x, m.
\]
Then we have
\[
\frac{d}{dt} H(x_t, m_t) \leq 0.
\]

**Proof.**
\[
\frac{d}{dt} H(x_t, m_t) = \nabla_x H^\top V_x + \nabla_m H^\top V_m
\]
\[
= (-V_x + aV_m)^\top V_x + (-V_m - aV_x)^\top V_m
\]
\[
= -(\hat{V}_x^\top V_x + \hat{V}_m^\top V_m) \leq 0.
\]

**Lemma B.5.** Under the condition of Theorem 3.1, let
\[
V_x(x, m) = \nabla K(\hat{m}) - \lambda x
\]
\[
V_m(x, m) = -\alpha \nabla f(x) - \gamma m = \frac{\hat{m} - m}{\varepsilon}
\]
and related
\[
\hat{V}_x(x, m) = \hat{m} - \nabla K^*(\lambda x) = -\varepsilon \alpha \nabla f(x) + (1 - \varepsilon \gamma)m - \nabla K^*(\lambda x),
\]
\[
\hat{V}_m(x, m) = \nabla K(\hat{m}) - \nabla K(m).
\]
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,
\[
\nabla_x H(x, m) = -\eta' \hat{V}_x - \eta V_m
\]
\[
\nabla_m H(x, m) = -\eta' \hat{V}_m + \eta V_x,
\]
where \( \eta = \frac{1 - \varepsilon \gamma}{1 + \varepsilon \lambda} \) and \( \eta' = \frac{\gamma + \lambda}{1 + \varepsilon \lambda} \). This yields
\[
\frac{d}{dt} H(x_t, m_t) = \nabla_x H^\top V_x + \nabla_m H^\top V_m = - (\eta' \hat{V}_x^\top V_x + \eta \hat{V}_m^\top V_m) \leq 0.
\]

**Proof.** Let \( \eta = \frac{1 - \varepsilon \gamma}{1 + \varepsilon \lambda} \). We have We have
\[
\nabla_m H(x, m) = \eta (\nabla K(m) - \lambda x)
\]
\[
= \eta (\nabla K(\hat{m}) - \lambda x + \nabla K(m) - \nabla K(\hat{m}))
\]
\[
= (\hat{V}_x - \hat{V}_m).
\]
\[
\nabla_x H(x, m)
\]
\[
= \alpha \nabla \nabla f(x) + \gamma \nabla K^*(\lambda x) + \eta (\lambda \nabla K^*(\lambda x) - \lambda m)
\]
\[
= \alpha \nabla f(x) + (\gamma + \eta \lambda) \nabla K^*(\lambda x) - \eta \lambda m
\]
\[
= (\gamma + \eta \lambda)(\varepsilon \alpha \nabla f(x) - (1 - \varepsilon \gamma)m + \nabla K^*(\lambda x)) + (\alpha - (\gamma + \eta \lambda)\varepsilon \alpha) \nabla f(x) - (\eta \lambda - (\gamma + \eta \lambda)(1 - \varepsilon \gamma))m
\]
\[
= \frac{\gamma + \lambda}{1 + \varepsilon \lambda} (\varepsilon \alpha \nabla f(x) - (1 - \varepsilon \gamma)m + \nabla K^*(\lambda x)) + \eta \alpha \nabla f(x) + \eta \gamma m
\]
\[
= -\gamma + \frac{\lambda}{1 + \varepsilon \lambda} \hat{V}_x - \eta V_m,
\]
where we used the following identities on \( \eta \):
\[
(\gamma + \eta \lambda) = \gamma + \frac{1 - \varepsilon \gamma}{1 + \varepsilon \lambda} \lambda \frac{\gamma + \lambda}{1 + \varepsilon \lambda} = \frac{\gamma + \lambda}{1 + \varepsilon \lambda}
\]
\[
1 - (\gamma + \eta \lambda)\varepsilon = 1 - \gamma + \frac{\lambda}{1 + \varepsilon \lambda} \varepsilon = \frac{1 - \varepsilon \gamma}{1 + \varepsilon \lambda} = \eta
\]
\[
\eta \lambda - (\gamma + \eta \lambda)(1 - \varepsilon \gamma) = -\gamma + \frac{\lambda}{1 + \varepsilon \lambda} \varepsilon \gamma = \frac{\varepsilon \gamma^2 - \gamma}{1 + \varepsilon \lambda} = -\gamma \eta.
\]

\[\square\]
B.9 Constraint Enforcing: Continuous Time

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

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

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

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

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

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

$$\lambda x_t = (1 - w_{s-t})z_{s-t} + w_{s-t}(\lambda x_s), \quad \text{where} \quad z_{s-t} = \frac{\int_t^s w_{t'} \nabla K(\tilde{m}_{t'}) dt'}{\int_t^s w_{t'} dt'}, \quad \forall 0 \leq s \leq t.$$

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

$$\text{dist}(\lambda x_t, \text{dom} K^*) = \inf_{z \in \text{dom} K^*} \| \lambda x_t - z \| \leq \| \lambda x_t - (1 - w_{s-t})z_{s-t} - w_{s-t}(\lambda \tilde{x}_s) \|$$

$$= w_{s-t} \| \lambda x_s - \lambda \tilde{x}_s \|$$

$$\leq \exp(\lambda(s-t))(\text{dist}(\lambda x_s, \text{dom} K^*) + \epsilon).$$

Taking $\epsilon \to 0$ yields

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

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

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

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

B.10 Discrete Time Analysis

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

$$m_{t+1} = \beta_2 m_t - (1 - \beta_2) \nabla f(x_t)$$

$$\tilde{m}_{t+1} = \beta_1 m_t - (1 - \beta_1) \nabla f(x_t)$$

$$x_{t+1} = x_t + \epsilon (\nabla K(m_{t+1}) - \lambda x_{t+1}),$$

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

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

and

$$\Delta_1^t = (\nabla K(\tilde{m}_{t+1}) - \lambda x_{t+1})^T (\tilde{m}_{t+1} - \nabla K^*(\lambda x_{t+1})), $$

$$\Delta_2^t = (\nabla K(\tilde{m}_{t+1}) - \nabla K(m_{t+1}))^T (\tilde{m}_{t+1} - m_{t+1}).$$

26
where $\nabla K^*$ is a subgradient of $K^*$. Then we have $\Delta_1^t \geq 0$ and $\Delta_2^t \geq 0$ from Lemma B.9, and
\[
H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \leq -\epsilon(a\Delta_1^t + b\Delta_2^t) + \frac{L\epsilon^2}{2}\|\nabla K(\tilde{m}_{t+1}) - \lambda x_{t+1}\|^2,
\]
where
\[
a = \frac{\epsilon \lambda \beta_1}{\epsilon \lambda (1 - \beta_1) + (1 - \beta_2)} + 1 \geq 0, \quad b = \frac{\beta_1 (1 - \beta_2)}{(\beta_2 - \beta_1)(\epsilon \lambda (1 - \beta_1) + (1 - \beta_2))} \geq 0.
\]
Hence, a telescoping sum yields
\[
\frac{1}{T} T \sum_{t=0}^{T-1} a\Delta_1^t + b\Delta_2^t \leq \frac{H(x_0, m_0) - H(x_T, m_T)}{\epsilon T} + \frac{L\epsilon}{2} B_t,
\]
where $B_t = \frac{1}{T} \sum_{t=1}^{T} \|\nabla K(\tilde{m}_{t+1}) - \lambda x_{t+1}\|^2$.

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

**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
\[
(\nabla K(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top (\nabla K^*(\lambda x_{t+1}) - \tilde{m}_{t+1}) \leq 0. \tag{38}
\]
\[
(\nabla K(\tilde{m}_{t+1}) - \nabla K(m_{t+1}))^\top (m_{t+1} - \tilde{m}_{t+1}) \leq 0. \tag{39}
\]
Taking $a \times$ Eq.(38) $+ b \times$ Eq.(39) for $a, b \geq 0$, we have
\[
(\nabla K(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top (a(\nabla K^*(\lambda x_{t+1}) - \tilde{m}_{t+1}) + b(m_{t+1} - \tilde{m}_{t+1})) + \cdots
\]
\[
+ b(\nabla K(m_{t+1}) - \lambda x_{t+1})^\top (-m_{t+1} + \tilde{m}_{t+1}) \leq 0.
\]

Plugging (43) yields
\[
(\nabla K(\tilde{m}_{t+1}) - \lambda x_{t+1})^\top (a\nabla K^*(\lambda x_{t+1}) - ((a + b)\beta_1 - b\beta_2)m_t + (a - (a + b)\beta_1 + b\beta_2)\nabla f(x_t))
\]
\[
- b(\beta_2 - \beta_1)(\nabla K(m_{t+1}) - \lambda x_{t+1})^\top (m_{t+1} + \nabla f(x_t)) \leq 0
\]

Define
\[
H(x, m) = (a - \epsilon) f(x) + \frac{a}{\lambda} K^*(\lambda x) + \frac{c}{\lambda} K(m) - cx^\top m, \quad \text{with} \quad c = (a + b)\beta_1 - b\beta_2,
\]
and
\[
\tilde{H}_t = (a - \epsilon) f(x_t) + \frac{a}{\lambda} K^*(\lambda x_{t+1}) - c m_t, \quad \hat{H}_t = \frac{c}{\lambda} \nabla K(m_{t+1}) - cx_{t+1}.
\]
Then the inequality can be written into
\[
\tilde{H}_t^\top (\nabla K(\tilde{m}_{t+1}) - \lambda x_{t+1}) + \hat{H}_t^\top \left( \frac{b(\beta_2 - \beta_1)\lambda}{c} (-m_t - \nabla f(x_t)) \right) \leq 0.
\]

Plugging the update rule of $x_{t+1} = x_t + \epsilon(\nabla K(\tilde{m}_{t+1}) - \lambda x_{t+1})$ and $m_{t+1} - m_t = -(1 - \beta_2)(m_t + \nabla f(x_t))$, we get
\[
\tilde{H}_t^\top \left( \frac{x_{t+1} - x_t}{\epsilon} \right) + \hat{H}_t^\top \left( \frac{b(\beta_2 - \beta_1)\lambda}{c(1 - \beta_2)} (m_{t+1} - m_t) \right) \leq 0.
\]

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 \( a, b, c \):

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

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

\[
z(\beta_2 - \beta_1)\lambda = (z(\beta_2 - \beta_1)\lambda + 1 + z(1 - \beta_2))\beta_1 - z(1 - \beta_2)\beta_2.
\]

We get

\[
z = \frac{\beta_1}{\epsilon(\beta_2 - \beta_1)\lambda - \epsilon(\beta_2 - \beta_1)\beta_1(1 - \beta_2)\beta_1 + (1 - \beta_2)\beta_2}.
\]

\[
= \frac{\beta_1}{\epsilon\lambda(\beta_2 - \beta_1)(1 - \beta_1) + (1 - \beta_2)(\beta_2 - \beta_1)}.
\]

\[
\geq \frac{\beta_1}{(\beta_2 - \beta_1)(\epsilon\lambda(1 - \beta_1) + (1 - \beta_2))} \geq 0.
\]

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

\[
H(x, m) = f(x) + \frac{1}{\lambda}K^*(\lambda x) + c(K^*(\lambda x) + K(m) - \lambda x^T m)
\]

\[
= f(x) + \frac{1}{\lambda}K^*(\lambda x) + \frac{\epsilon\lambda\beta_1}{\epsilon\lambda(1 - \beta_1) + (1 - \beta_2)}(K^*(\lambda x) + K(m) - \lambda x^T m),
\]

and

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

From Lemma B.9, we get

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

\[\square\]

**Lemma B.9.** Let \( H(x, m) = f(x) + K_1(x) + K_2(m) - \lambda xm \), where \( f \) is \( L \)-smooth, and \( K_1, K_2 \) are convex functions with subgradient \( \nabla K_1 \) and \( \nabla K_2 \). Then

\[
H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \leq \hat{\nabla}_x H_t^T (x_{t+1} - x_t) + \hat{\nabla}_m H_t^T (m_{t+1} - m_t) + \frac{L}{2} \|x_{t+1} - x_t\|_2^2,
\]

where

\[
\hat{\nabla}_x H_t = \nabla f(x_t) + K_1(x_{t+1}) - \lambda m_t
\]

\[
\hat{\nabla}_m H_t = K_2(m_{t+1}) - \lambda x_{t+1}.
\]

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

\[
f(x_{t+1}) - f(x_t) \leq \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{L}{2} \|x_{t+1} - x_t\|_2^2
\]

\[
K_1(x_{t+1}) - K_1(x_t) \leq \nabla K_1(x_{t+1})^T (x_{t+1} - x_t)
\]

\[
K_2(m_{t+1}) - K_2(m_t) \leq \nabla K_2(m_{t+1})^T (m_{t+1} - m_t)
\]

\[
x_{t+1}^T m_{t+1} - x_t^T m_t = m_t^T (x_{t+1} - x_t) + x_{t+1}^T (m_{t+1} - m_t).
\]

Summing them together yields the result. \[\square\]
Theorem B.10. Under the same conditions of Theorem 4.1, for any two integers \( s \leq t \),
\[
\text{dist}(\lambda x_t, \text{dom}\mathcal{K}^*) \leq \left( \frac{1}{1 + \epsilon \lambda} \right)^{s-t} \text{dist}(\lambda x_s, \text{dom}\mathcal{K}^*), \quad \forall s \leq t.
\]

Proof. Rewriting the update into the explicit form:
\[
x_{t+1} = \frac{1}{1 + \epsilon \lambda} x_t + \frac{\epsilon}{1 + \epsilon \lambda} \nabla K(\tilde{m}_{t+1}).
\]
Unrolling this update yields, with \( w_{s \to t} = \left( \frac{1}{1 + \epsilon \lambda} \right)^{s-t} \),
\[
\lambda x_t = (1 - w_{s \to t}) z_{s \to t} + w_{s \to t} x_s, \quad z_{s \to t} = \frac{\sum_{k=s+1}^{t} w_{k \to t} \nabla K(\tilde{m}_k)}{\sum_{k=s+1}^{t} w_{k \to t}}.
\]
We have \( \nabla K(\tilde{m}_k) \in \text{dom}\mathcal{K}^* \) from Lemma B.7 and \( \text{dom}\mathcal{K}^* \) is convex. Hence \( z_{s \to t} \), as the convex combination of \( \{\nabla K(\tilde{m}_k)\}_k \), belongs to \( \text{dom}\mathcal{K}^* \). For any \( \epsilon > 0 \), let \( \lambda \hat{x}_s \in \text{dom}\mathcal{K}^* \) to the point satisfying \( \|\lambda \hat{x}_s - \lambda x_s\| \leq \text{dist}(\lambda x_s, \text{dom}\mathcal{K}^*) + \epsilon \). Hence,
\[
\text{dist}(\lambda x_t, \text{dom}\mathcal{K}^*) = \inf_{z \in \text{dom}\mathcal{K}^*} \|\lambda x_t - z\|
\leq \|\lambda x_t - (1 - w_{s \to t}) z_{s \to t} + w_{s \to t} \lambda \hat{x}_s\|
= w_{s \to t} \|\lambda x_s - \lambda \hat{x}_s\|
\leq \left( \frac{1}{1 + \epsilon \lambda} \right)^{s-t} (\text{dist}(\lambda x_s, \text{dom}\mathcal{K}^*) + \epsilon).
\]
Taking \( \epsilon \to 0 \) yields the result.

\[\Box\]

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 \( \Delta_1^t \) and \( \Delta_2^t \).

Compared with the deterministic case, the main challenge is to bound an additional correlation term due to the stochastic gradient at each iteration \( t \):
\[
V_t := \text{cov}(g_t, \nabla K(\tilde{m}_{t+1})) = \text{cov}(g_t, \nabla K(\beta_1 m_t + (1 - \beta_1) g_t)),
\]
where \( \text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])^\top (Y - \mathbb{E}[Y])] \).

Definition B.11. For a random variable \( X \) on \( \mathbb{R}^d \), its (trace of) variance \( \text{var}(X) \), when exists, is defined as
\[
\text{var}(X) = \mathbb{E}[\|X - \mathbb{E}[X]\|^2_2]
\]

Assumption B.12. Assume
\[
\text{var}(g_t) \leq \frac{v_{\text{max}}}{n_{\text{batch}}},
\]
where \( n_{\text{batch}} \) represents the batch size.

Assumption B.13. \( \mathcal{D} \) is the data distribution, the stochastic sample \( \xi_t \sim \mathcal{D} \) 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 \( f : \mathbb{R}^d \to \mathbb{R} \) that is \( L \)-smooth. Additionally, let \( \mathcal{K} : \mathbb{R}^d \to \mathbb{R} \) be a closed and convex function, consider the following scheme:
\[
\begin{align*}
    m_{t+1} &= \beta_2 m_t - (1 - \beta_2) g_t \\
    \tilde{m}_{t+1} &= \beta_1 m_t - (1 - \beta_1) g_t \\
    x_{t+1} &= x_t + \epsilon (\nabla \mathcal{K}(\tilde{m}_{t+1}) - \lambda x_{t+1}).
\end{align*}
\]
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 $E[g_t] = \nabla f(x_t)$. $\nabla K$ is a weak gradient of $K$ with $\mathbf{y} = 0$, $\|\nabla K(x) - \nabla K(y)\| \leq L_K \|x - y\|$, $x, y \in \mathbb{R}^d$, 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 $K$. Define the following Lyapunov function:

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

and

$$
\Delta^1_t = (\nabla K(m_{t+1}) - \lambda x_{t+1})^T (m_{t+1} - \mathcal{K}^*(\lambda x_{t+1})), \\
\Delta^2_t = (\nabla K(m_{t+1}) - \nabla K(m_{t+1}))^T (m_{t+1} - m_{t+1}),
$$

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

$$E[H(x_{t+1}, m_{t+1}) - H(x_t, m_t)] \leq E \left[ -\epsilon (a \Delta^1_t + b \Delta^2_t) + \frac{L e^2}{2} \|\nabla K(m_{t+1}) - \lambda x_{t+1}\|^2 \right] + e \frac{K}{1+\epsilon} (1-\beta_1) \frac{v_{\max}}{n_{\text{batch}}} + L_K \frac{1}{1+\epsilon} \sqrt{(1-\beta_2) \frac{v_{\max}}{n_{\text{batch}}}}$$

where

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

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

Hence, a telescoping sum yields

$$\frac{1}{T} \sum_{t=0}^{T-1} E[a \Delta^1_t + b \Delta^2_t] \leq E \left[ \frac{H(x_0, m_0) - H(x_T, m_T)}{e T} + \frac{L e}{2} B_t + \frac{C_t}{n_{\text{batch}}} \right],$$

where $B_t = \frac{1}{T} \sum_{t=1}^{T} \|\nabla K(m_{t+1}) - \lambda x_{t+1}\|^2$, and $C_t = \frac{L K}{1+\epsilon} (1-\beta_1) + L_K \frac{1}{1+\epsilon} \sqrt{(1-\beta_2) \frac{v_{\max}}{n_{\text{batch}}}}$.

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

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

where

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

By the definition of $\Delta^1_t, \Delta^2_t$, we have

$$a \Delta^1_t + b \Delta^2_t = a (\nabla K(m_{t+1}) - \lambda x_{t+1})^T (m_{t+1} - \mathcal{K}^*(\lambda x_{t+1})) + b (\nabla K(m_{t+1}) - \nabla K(m_{t+1}))^T (m_{t+1} - m_{t+1})$$

$$= (\nabla K(m_{t+1}) - \lambda x_{t+1})^T (a (\nabla K^*(\lambda x_{t+1}) - m_{t+1}) + b (m_{t+1} - m_{t+1})) + b (\nabla K(m_{t+1}) - \lambda x_{t+1})^T (m_{t+1} - m_{t+1})$$

$$= -((\nabla K(m_{t+1}) - \lambda x_{t+1})^T (a \nabla K^*(\lambda x_{t+1}) - m_{t+1} + (a - (a + b) \beta_1 + b \beta_2) m_t + (a - (a + b) \beta_1 + b \beta_2) \nabla f(x_t)))$$

$$- b \beta_2 - b \beta_1 \frac{c}{1-\beta_2} \cdot \frac{c}{\lambda} \nabla K(m_{t+1}) - c x_{t+1})^T (m_{t+1} - m_t)$$

$$= - [(a - c) g_t + a \nabla K^*(\lambda x_{t+1}) - cm_t]^T (\nabla K(m_{t+1}) - \lambda x_{t+1})$$

$$- \frac{1}{\epsilon} \left[ \frac{c}{\lambda} \nabla K(m_{t+1}) - c x_{t+1} \right]^T (m_{t+1} - m_t)$$

$$= - \frac{1}{\epsilon} [(a - c) g_t + a \nabla K^*(\lambda x_{t+1}) - cm_t]^T (x_{t+1} - x_t)$$

$$- \frac{1}{\epsilon} \left[ \frac{c}{\lambda} \nabla K(m_{t+1}) - c x_{t+1} \right]^T (m_{t+1} - m_t)$$ (42)
By Lemma B.9,
\[ H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \leq \nabla_x H^T_t (x_{t+1} - x_t) + \nabla_m H^T_t (m_{t+1} - m_t) + \frac{L}{2} \| x_{t+1} - x_t \|^2, \]
where
\[ \nabla_x H_t = (a - c)\nabla f(x_t) + a \nabla K^* (\lambda x_{t+1}) - cm_t, \]
\[ \nabla_m H_t = \frac{c}{\lambda} \nabla K (m_{t+1}) - cx_{t+1} = \frac{c}{\lambda} (\nabla x_t - \nabla K (\tilde{m}_{t+1}) + \nabla K (m_{t+1})) \]
with
\[ V_{x,t} = x_{t+1} - x_t = c (\nabla K (\tilde{m}_{t+1}) - \lambda x_{t+1}) \]
\[ V_{m,t} = m_{t+1} - m_t = -(1 - \beta_2)(g_t - m_t) \]
\[ \tilde{m}_{t+1} - m_{t+1} = - (\beta_2 - \beta_1)(g_t - m_t) = - (\beta_2 - \beta_1) V_{m,t} \]
\[ \tilde{V}_{m,t} = - \nabla K (\tilde{m}_{t+1}) + \nabla K (m_{t+1}) \]
This gives
\[ H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \leq \nabla_x H^T_t (x_{t+1} - x_t) + \nabla_m H^T_t (m_{t+1} - m_t) + \frac{L}{2} \| x_{t+1} - x_t \|^2 \]
Hence,
\[ H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \leq [(a - c)\nabla f(x_t) + a \nabla K^* (\lambda x_{t+1}) - cm_t]^T (x_{t+1} - x_t) \]
\[ + \left[ \frac{c}{\lambda} \nabla K (m_{t+1}) - cx_{t+1} \right]^T (m_{t+1} - m_t) + \frac{L}{2} \| x_{t+1} - x_t \|^2 \]
\[ = [(a - c)g_t + a \nabla K^* (\lambda x_{t+1}) - cm_t]^T (x_{t+1} - x_t) \]
\[ + \left[ \frac{c}{\lambda} \nabla K (m_{t+1}) - cx_{t+1} \right]^T (m_{t+1} - m_t) + \frac{L}{2} \| x_{t+1} - x_t \|^2 \]
\[ + \epsilon (a - c)\nabla f(x_t) - g_t \nabla K (\tilde{m}_{t+1}) - \lambda x_{t+1} \]
\[ = - \epsilon (a \Delta^1_t + b \Delta^2_t) + \frac{L}{2} \| x_{t+1} - x_t \|^2 \quad \text{by equation 44} \]
\[ + \epsilon (a - c)\nabla f(x_t) - g_t \nabla K (\tilde{m}_{t+1}) - \lambda x_{t+1} \]
It suffices to bound \( \mathbb{E} [(\nabla f(x_t) - g_t)^T (\nabla K (\tilde{m}_{t+1}) - \lambda x_{t+1})] \).
Note that
\[ \mathbb{E} [(\nabla f(x_t) - g_t)^T (\nabla K (\tilde{m}_{t+1}) - \lambda x_{t+1})] \]
\[ = \mathbb{E} \left[ (\nabla f(x_t) - g_t)^T \left( \frac{1}{1 + \lambda \epsilon} \nabla K (\tilde{m}_{t+1}) - \frac{\lambda}{1 + \lambda \epsilon} x_t \right) \right] \]
\[ = \frac{1}{1 + \lambda \epsilon} \mathbb{E} \left[ (\nabla f(x_t) - g_t)^T \nabla K (\tilde{m}_{t+1}) \right] + \frac{\lambda}{1 + \lambda \epsilon} \mathbb{E} \left[ (\nabla f(x_t) - g_t)^T x_t \right] \]
By Assumption B.13,
\[ \mathbb{E} \left[ (\nabla f(x_t) - g_t)^T \lambda x_t \right] = \lambda \mathbb{E}_{x_t} \left[ (\nabla f(x_t) - \nabla f(x_t, \xi_t))^T x_t \mid x_t \right] \]
\[ = 0 \quad \text{by B.13} \quad \mathbb{E}[\nabla f(x, \xi)] = \nabla f(x) \]
Next, let us bound \( \mathbb{E} [(\nabla f(x_t) - g_t)^T \nabla K (\tilde{m}_{t+1})] \).
\[ \mathbb{E} \left[ (\nabla f(x_t) - g_t)^T \nabla K (\tilde{m}_{t+1}) \right] = \mathbb{E} \left[ (\nabla f(x_t) - g_t)^T \nabla K (\beta_1 m_t - (1 - \beta_1) g_t) \right] \]
\[ \leq L_K (1 - \beta_1) \var{g_t} + L_K \sqrt{\var{m_t} \cdot \var{g_t}} \quad \text{by B.18} \]
\[ \leq L_K (1 - \beta_1) \frac{v_{\max}}{n_{\text{batch}}} + L_K \sqrt{\frac{1 - \beta_2}{1 + \beta_2} \frac{v_{\max}}{n_{\text{batch}}} } \quad \text{by B.18} \]
Lemma B.15. Let $X, Y$ be two $\mathbb{R}^d$-valued random variables with $\text{var}(X) < +\infty$ and $\text{var}(Y) < +\infty$, and assume $\mathcal{K}$ yields a weak derivative $\nabla \mathcal{K}$. We have

$$\mathbb{E}[(Y - \mathbb{E}[Y])^\top \nabla \mathcal{K}(X + c Y)] \leq L_K \text{var}(Y) + L_K \text{var}(X) \cdot \text{var}(Y)$$

Proof.

$$\mathbb{E}[(Y - \mathbb{E}[Y])^\top \nabla \mathcal{K}(X + c Y)] = \mathbb{E}[(Y - \mathbb{E}[Y])^\top (\nabla \mathcal{K}(X + c Y) - \mathcal{K}(\mathbb{E}[X] + c \mathbb{E}[Y]))]$$

$$= \sqrt{\mathbb{E}[(Y - \mathbb{E}[Y])^2]} \sqrt{\mathbb{E}[\nabla \mathcal{K}(X + c Y) - \mathcal{K}(\mathbb{E}[X] + c \mathbb{E}[Y])]^2}$$

$$= \sqrt{\mathbb{E}[(Y - \mathbb{E}[Y])^2]} L_K \mathbb{E}[|X + c Y - \mathbb{E}[X] - c \mathbb{E}[Y]|]^2$$

$$= L_K \sqrt{\mathbb{E}[(Y - \mathbb{E}[Y])^2]} (\sqrt{\mathbb{E}[|X - \mathbb{E}[X]|^2]} + \sqrt{\mathbb{E}[|Y - \mathbb{E}[Y]|^2]})$$

$$= L_K \mathbb{E}[|Y - \mathbb{E}[Y]|^2] + L_K \sqrt{\mathbb{E}[(Y - \mathbb{E}[Y])^2]} \sqrt{\mathbb{E}[|X - \mathbb{E}[X]|^2]}$$

$$= L_K \text{var}(Y) + L_K \sqrt{\text{var}(X)} \cdot \text{var}(Y)$$

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

$$\mathbb{E}\left[\left(\sum_{t=1}^k \alpha_t \delta_t\right)^2\right] \leq \sum_{t=1}^k \alpha_t^2 \sigma^2.$$

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

$$\mathbb{E}[|Y_k|] = \mathbb{E}\left[\left|\sum_{t=1}^k \alpha_t \delta_t\right|\right]$$

$$\leq \sum_t |\alpha_t| \mathbb{E}[|\delta_t|] \quad \text{triangle inequality}$$

$$= \sum_t |\alpha_t| \mathbb{E}[\mathbb{E}[|\delta_t||x_t]] \quad \text{law of total probability}$$

$$\leq \sum_t |\alpha_t| \mathbb{E}[\sqrt{\mathbb{E}[\delta_t^2|x_t]}] \quad \text{Jensen’s inequality}$$

$$\leq \sum_t |\alpha_t| \sigma < \infty$$
Second, again using the law of total probability,
\[
\mathbb{E}[Y_{k+1}|Y_1, \ldots, Y_k] = \mathbb{E} \left[ \sum_{l=1}^{k+1} \alpha_l \delta_l \right]
\]
\[
= Y_k + \alpha_{k+1} \mathbb{E} \left[ \delta_{k+1}|\alpha_1 \delta_1, \ldots, \alpha_k \delta_k \right]
\]
\[
= Y_k + \alpha_{k+1} \mathbb{E} \left[ \mathbb{E} \left[ \delta_{k+1}|x_{k+1}, \alpha_1 \delta_1, \ldots, \alpha_k \delta_k \right]|\alpha_1 \delta_1, \ldots, \alpha_k \delta_k \right]
\]
\[
= Y_k + \alpha_{k+1} \mathbb{E} \left[ \mathbb{E} \left[ \delta_{k+1}|x_{k+1} \right]|\alpha_1 \delta_1, \ldots, \alpha_k \delta_k \right]
\]
\[
= Y_k
\]
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.
\[
\mathbb{E}[\sum_{l=1}^{k} \alpha_l \delta_l^2] = \sum_{l=1}^{k} \mathbb{E}[\alpha_l \delta_l^2] + 2 \sum_{l<j} \mathbb{E}[\alpha_l \alpha_j \delta_l \delta_j]
\]
\[
= \sum_{l=1}^{k} \alpha_l^2 \mathbb{E}[\delta_l^2|\delta_1, \ldots, \delta_{l-1}] + 2 \sum_{l<j} \alpha_l \alpha_j \mathbb{E}[\delta_l \mathbb{E}[\delta_j|\delta_1, \ldots, \delta_{j-1}]|\delta_l]
\]
\[
= \sum_{l=1}^{k} \alpha_l^2 \mathbb{E}[\mathbb{E}[\delta_l^2|x_l, \delta_1, \ldots, \delta_{l-1}]|\delta_1, \ldots, \delta_{l-1}] + 0
\]
\[
= \sum_{l=1}^{k} \alpha_l^2 \var \delta_l.
\]
\[\square\]

The consequence of this lemma is that we are able to treat \(\delta_1, \ldots, \delta_k\) as if they are independent, even though they are not—clearly \(\delta_l\) is dependent on \(\delta_1, \ldots, \delta_{l-1}\) through \(x_l\). By Lemma B.16, we can compute the variance of momentum \(m_t\),
\[
\text{var}(m_t) = (1 - \beta_2)^2 \mathbb{E} \left[ \sum_{i=1}^{t} \beta_2^{t-i} \delta_i \right]^2
\]
\[
= (1 - \beta_2)^2 \mathbb{E} \sum_{i=1}^{t} \beta_2^{2t-2i} \|\delta_i\|^2
\]
\[
= \frac{(1 - \beta_2)v_{\text{max}}}{(1 + \beta_2)^{n_{\text{batch}}}}
\]

B.12 Analysis with Stochastic Gradient LION

**Theorem B.17.** Under the assumptions delineated in B.13 and B.12, consider a function \(f : \mathbb{R}^d \rightarrow \mathbb{R}\) that is \(L\)-smooth. Consider the following scheme:
\[
m_{t+1} = \beta_2 m_t - (1 - \beta_2) g_t
\]
\[
m^t_{t+1} = \beta_1 m_t - (1 - \beta_1) g_t
\]
\[
x_{t+1} = x_t + \epsilon (\text{sign}(m^t_{t+1}) - \lambda x_{t+1}),
\]
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), \text{ and } \beta_2 > \beta_1, \text{ and } \epsilon, \lambda > 0.
\]

Define the following Lyapunov function:
\[
H(x, m) = f(x) + \frac{1}{\lambda} \|\lambda x\|^* + \frac{\beta_1}{\epsilon(1 - \beta_1) + (1 - \beta_2)} (\|\lambda x\|^* + \|m\| - \lambda x^\top m),
\]
and
\[
\Delta^1_t = (\text{sign}(m_{t+1}) - \lambda x_{t+1})^\top (m^t_{t+1} - \text{sign}^*(\lambda x_{t+1})),
\]
\[
\Delta^2_t = (\text{sign}(m^t_{t+1}) - \text{sign}^*(m_{t+1}))^\top (\hat{m}_{t+1} - \hat{m}_{t+1}).
\]
where \(\text{sign}^*\) is a subgradient of \(K^*\). Then we have \(\Delta_1^t \geq 0\) and \(\Delta_2^t \geq 0\) from Lemma B.9, and

\[
\mathbb{E} [H(x_{t+1}, m_{t+1}) - H(x_t, m_t)] \leq \mathbb{E} \left[ -\epsilon (a\Delta_1^t + b\Delta_2^t) + \frac{L\epsilon^2}{2} \|\text{sign}(\hat{m}_{t+1}) - \lambda x_{t+1}\|^2_2 \right] + \epsilon \frac{1}{1 + \lambda \epsilon} \sqrt{d \cdot v_{\text{max}} + \frac{m_{\text{batch}}}{T}}
\]

where

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

\(v_{\text{max}}, m_{\text{batch}}\) are defined in B.12

Hence, a telescoping sum yields

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [a\Delta_1^t + b\Delta_2^t] \leq \mathbb{E} \left[ \frac{H(x_0, m_0) - H(x_T, m_T)}{\epsilon T} + \frac{L\epsilon}{2} B_t + \frac{1}{1 + \lambda \epsilon} \sqrt{d \cdot v_{\text{max}} + \frac{m_{\text{batch}}}{T}} \right],
\]

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

Proof. Define

\[
H(x, m) = (a - c) f(x) + \frac{a}{\lambda} \|\lambda x\|^2 + \frac{c}{\lambda} \|m\| - cx^T m, \quad \text{with} \quad c = (a + b)\beta_1 - b\beta_2,
\]

where

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

By the definition of \(\Delta_1^t, \Delta_2^t\), we have

\[
\begin{align*}
a \Delta_1^t + b \Delta_2^t &= \mathbb{E} [H(x_{t+1}, m_{t+1}) - H(x_t, m_t)] \\
&= (a \text{sign}(\hat{m}_{t+1}) - \lambda x_{t+1})^T (\hat{m}_{t+1} - \text{sign}^*(\lambda x_{t+1})) \\
&\quad + b (\text{sign}(\hat{m}_{t+1}) - \lambda x_{t+1})^T (\hat{m}_{t+1} - m_{t+1}) \\
&= (\text{sign}(\hat{m}_{t+1}) - \lambda x_{t+1})^T (a \text{sign}^*(\lambda x_{t+1}) - \text{sign}^*(\lambda m_{t+1})) \\
&\quad + b (\text{sign}(\hat{m}_{t+1}) - \lambda x_{t+1})^T (m_{t+1} - \hat{m}_{t+1}) \\
&= -b \frac{\beta_2 - \beta_1}{1 - \beta_2} \lambda \left( \frac{c}{\lambda} \text{sign}(m_{t+1}) - cx_{t+1}^T \right)^T (m_{t+1} - m_t) \\
&= -\left[ (a - c) g_t + \text{sign}^*(\lambda x_{t+1}) - cm_t \right]^T \left( \text{sign}(\hat{m}_{t+1}) - \lambda x_{t+1} \right) \\
&\quad - \frac{1}{\epsilon} \left[ \frac{c}{\lambda} \text{sign}(m_{t+1}) - cx_{t+1}^T \right]^T \left( m_{t+1} - m_t \right) \\
&= -\frac{1}{\epsilon} \left[ (a - c) g_t + \text{sign}^*(\lambda x_{t+1}) - cm_t \right]^T \left( x_{t+1} - x_t \right) \\
&\quad - \frac{1}{\epsilon} \left[ \frac{c}{\lambda} \text{sign}(m_{t+1}) - cx_{t+1}^T \right]^T \left( m_{t+1} - m_t \right)
\end{align*}
\]

By Lemma B.9,

\[
H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \leq \nabla_x H_t^T (x_{t+1} - x_t) + \nabla_m H_t^T (m_{t+1} - m_t) + \frac{L}{2} \|x_{t+1} - x_t\|^2_2,
\]

where

\[
\nabla_x H_t = (a - c) \nabla f(x_t) + \text{sign}^*(\lambda x_{t+1}) - cm_t,
\]

\[
\nabla_m H_t = \frac{c}{\lambda} \text{sign}(m_{t+1}) - cx_{t+1}^T = \frac{c}{\epsilon \lambda} (\hat{V}_{x,t} - \text{sign}(\hat{m}_{t+1}) + \text{sign}(m_{t+1}))
\]

34
Hence,

This gives

\[
H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \\
\leq \hat{\nabla}_z H^T_t (x_{t+1} - x_t) + \hat{\nabla}_m H^T_t (m_{t+1} - m_t) + \frac{L}{2} \| x_{t+1} - x_t \|_2^2
\]

Hence,

\[
H(x_{t+1}, m_{t+1}) - H(x_t, m_t) \leq [(a - c) \nabla f(x_t) + \text{asign}(\lambda x_{t+1}) - cm_t]^T (x_{t+1} - x_t)
\]
\[
+ \left[ \frac{c}{\lambda} \text{sign}(m_{t+1}) - c x_{t+1} \right]^T (m_{t+1} - m_t) + \frac{L}{2} \| x_{t+1} - x_t \|_2^2
\]
\[
= [(a - c) g_t + \text{asign}(\lambda x_{t+1}) - cm_t]^T (x_{t+1} - x_t)
\]
\[
+ \left[ \frac{c}{\lambda} \text{sign}(m_{t+1}) - c x_{t+1} \right]^T (m_{t+1} - m_t) + \frac{L}{2} \| x_{t+1} - x_t \|_2^2
\]
\[
+ \epsilon (a - c) (\nabla f(x_t) - g_t)^T (\text{sign}(\tilde{m}_{t+1}) - \lambda x_{t+1})
\]
\[
= -\epsilon (a \Delta^1_t + b \Delta^2_t) + \frac{L}{2} \| x_{t+1} - x_t \|_2^2 \quad \text{by equation 44}
\]
\[
+ \epsilon (a - c) (\nabla f(x_t) - g_t)^T (\text{sign}(\tilde{m}_{t+1}) - \lambda x_{t+1})
\]

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

Note that

\[
\mathbb{E} \left[ (\nabla f(x_t) - g_t)^T (\text{sign}(\tilde{m}_{t+1}) - \lambda x_{t+1}) \right]
\]
\[
= \mathbb{E} \left[ (\nabla f(x_t) - g_t)^T \left( \frac{1}{1 + \lambda \epsilon} \text{sign}(\tilde{m}_{t+1}) - \frac{\lambda}{1 + \lambda \epsilon} x_t \right) \right]
\]
\[
= \frac{1}{1 + \lambda \epsilon} \mathbb{E} \left[ (\nabla f(x_t) - g_t)^T \text{sign}(\tilde{m}_{t+1}) \right] + \frac{\lambda}{1 + \lambda \epsilon} \mathbb{E} \left[ \text{sign}(\tilde{m}_{t+1}) \right]
\]

By Assumption B.13,

\[
\mathbb{E} \left[ (\nabla f(x_t) - g_t)^T \lambda x_t \right] = \lambda \mathbb{E}_{x_t} \left[ \mathbb{E}_{\xi_t} \left[ (\nabla f(x_t) - \nabla f(x_t, \xi_t))^T x_t \mid x_t \right] \right]
\]
\[
= 0 \quad \text{by B.13} \quad \mathbb{E}[\nabla f(x, \xi)] = \nabla f(x)
\]

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

\[
\mathbb{E} \left[ (\nabla f(x_t) - g_t)^T \text{sign}(\tilde{m}_{t+1}) \right] = \mathbb{E} \left[ (\nabla f(x_t) - g_t)^T \text{sign}(\beta_1 m_t - (1 - \beta_1) g_t) \right]
\]
\[
\leq \sqrt{d \cdot \text{var}(g_t)} \quad \text{by B.18}
\]
\[
\leq \sqrt{d \cdot v_{\max} / n_{\text{batch}}} \quad \text{by B.12}
\]

Hence,

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

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

Proof.

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