Organization of the Appendix

A Proofs for Section 3

B Proofs for Section 4

C Experimental details

A Proofs for Stochastic Composite Minimization

A.1 Proof of Proposition 3

We break down the proof of Proposition 3 into different lemmas. The first one bounds the improvement on G over an iteration, the second one bounds the improvement H over an iteration, the third one combines the first two results and further exploits the structure of the functions at play. The proof of the proposition then follows by carefully choosing A_{k+1} as a function of A_k. In the remainder of the section, we work with the assumptions stated in Proposition 3 and do not restate them in the statements of the lemmas.

Lemma 1.

$$A_{k+1}G(y_{k+1}) - c_{k+1} \leq A_k G(y_k) - c_k + (A_{k+1} - A_k)\langle \nabla G(v_k),\, z_{k+1}\rangle + \frac{\beta}{2}\left(A_{k+1} - 2A_k + \frac{A_k^2}{A_{k+1}}\right)\|z_{k+1} - z_k\|^2 .$$

Proof. By smoothness

$$\begin{aligned}
A_{k+1}G(y_{k+1}) &\leq A_{k+1}\left(G(v_k) + \langle \nabla G(v_k),\, y_{k+1} - v_k\rangle + \frac{\beta}{2}\|y_{k+1} - v_k\|^2\right)\\
&= \underbrace{(A_{k+1} - A_k)\left(G(v_k) - \langle \nabla G(v_k),\, v_k\rangle\right)}_{=c_{k+1}-c_k} + A_k\left(G(v_k) - \langle \nabla G(v_k),\, v_k\rangle\right)\\
&\quad + A_{k+1}\langle \nabla G(v_k),\, y_{k+1}\rangle + A_{k+1}\frac{\beta}{2}\|y_{k+1} - v_k\|^2\\
&= c_{k+1} - c_k + A_k\left(G(v_k) + \langle \nabla G(v_k),\, y_k - v_k\rangle\right) - A_k\langle \nabla G(v_k),\, y_k\rangle\\
&\quad + A_{k+1}\langle \nabla G(v_k),\, y_{k+1}\rangle + A_{k+1}\frac{\beta}{2}\|y_{k+1} - v_k\|^2\\
&\leq c_{k+1} - c_k + A_k G(y_k) - A_k\langle \nabla G(v_k),\, y_k\rangle\\
&\quad + A_{k+1}\langle \nabla G(v_k),\, y_{k+1}\rangle + A_{k+1}\frac{\beta}{2}\|y_{k+1} - v_k\|^2
\end{aligned}$$

where the last inequality follows by strong convexity of G, Now, since $y_{k+1} = \frac{A_k}{A_{k+1}}y_k + \frac{A_{k+1}-A_k}{A_{k+1}}z_{k+1}$, we can simplify to

$$A_{k+1}G(y_{k+1}) \leq c_{k+1} - c_k + A_k G(y_k) + (A_{k+1} - A_k)\langle \nabla G(v_k),\, z_{k+1}\rangle + A_{k+1}\frac{\beta}{2}\|y_{k+1} - v_k\|^2 .$$

Finally,

$$A_{k+1}\|y_{k+1} - v_k\|^2 = A_{k+1}\tau_k^2\|z_k - z_{k+1}\|^2 = \frac{(A_{k+1} - A_k)^2}{A_{k+1}}\|z_{k+1} - z_k\|^2$$

which concludes the proof. □

Lemma 2. *For any $g_H(z_k) \in \partial H(z_k)$,*

$$A_{k+1}H(y_{k+1}) - A_{k+1}H(z_{k+1}) \leq A_k H(y_k) - A_k H(z_k) - A_k\langle g_H(z_k),\, z_{k+1} - z_k\rangle - \frac{\mu}{2}A_k\left(1 - \frac{A_k}{A_{k+1}}\right)\|z_{k+1} - y_k\|^2 - A_k\frac{\mu}{2}\|z_{k+1} - z_k\|^2 .$$

Proof. By strong convexity of H, for any $g_H(y_{k+1}) \in \partial H(y_{k+1})$ we have

$$\begin{aligned} A_{k+1}H(y_{k+1}) &= (A_{k+1} - A_k)\, H(y_{k+1}) + A_k H(y_{k+1}) \\ &\leq (A_{k+1} - A_k) \left(H(z_{k+1}) - \langle g_H(y_{k+1}),\, z_{k+1} - y_{k+1} \rangle - \frac{\mu}{2} \left\| z_{k+1} - y_{k+1} \right\|^2 \right) \\ &\quad + A_k \left(H(y_k) - \langle g_H(y_{k+1}),\, y_k - y_{k+1} \rangle - \frac{\mu}{2} \left\| y_k - y_{k+1} \right\|^2 \right). \end{aligned}$$

Now observe that

$$\begin{aligned} z_{k+1} - y_{k+1} &= (1 - \tau_k)(z_{k+1} - y_k) = \frac{A_k}{A_{k+1}}(z_{k+1} - y_k) \\ y_k - y_{k+1} &= \tau_k(y_k - z_{k+1}) = -\frac{A_{k+1} - A_k}{A_{k+1}}(z_{k+1} - y_k) \end{aligned}$$

and thus we see that the two inner product terms cancel out. Moreover we can also simplify the norms and get

$$A_{k+1}H(y_{k+1}) \leq A_k H(y_k) + (A_{k+1} - A_k)H(z_{k+1}) - \frac{\mu}{2} \left\| z_{k+1} - y_k \right\|^2 \left(\underbrace{(A_{k+1} - A_k)\tfrac{A_k^2}{A_{k+1}^2} + A_k \tfrac{(A_{k+1} - A_k)^2}{A_{k+1}^2}}_{= A_k - \frac{A_k^2}{A_{k+1}}} \right).$$

Finally, observe that by strong convexity of H again, for any $g_H(z_k) \in \partial H(z_k)$,

$$-A_k H(z_{k+1}) \leq -A_k H(z_k) - A_k \langle g_H(z_k),\, z_{k+1} - z_k \rangle - A_k \frac{\mu}{2} \left\| z_{k+1} - z_k \right\|^2,$$

which concludes the proof. □

Lemma 3. *Defining $m_k(x) = \langle d_k,\, x \rangle + c_k + A_k H(x) + \beta w(x)$, we have*

$$\begin{aligned} A_{k+1}F(y_{k+1}) - m_{k+1}(z_{k+1}) \leq\, & A_k F(y_k) - m_k(z_k) \\ & + (A_{k+1} - A_k)\langle \nabla G(v_k) - g_k,\, z_{k+1} \rangle \\ & - \frac{1}{2} \left(A_k(\mu + 2\beta) + \beta(\nu - A_{k+1}) - \beta \frac{A_k^2}{A_{k+1}} \right) \left\| z_{k+1} - z_k \right\|^2 \\ & - \frac{\mu}{2} A_k \left(1 - \frac{A_k}{A_{k+1}} \right) \left\| z_{k+1} - y_k \right\|^2. \end{aligned}$$

Proof. Summing the inequalities in the above two lemmas above we have that for any $g_H(z_k) \in \partial H(z_k)$,

$$\begin{aligned} A_{k+1}F(y_{k+1}) - c_{k+1} - A_{k+1}H(z_{k+1}) \leq\, & A_k F(y_k) - c_k - A_k H(z_k) \\ & + (A_{k+1} - A_k)\langle \nabla G(v_k),\, z_{k+1} \rangle - A_k \langle g_H(z_k),\, z_{k+1} - z_k \rangle \\ & - \frac{1}{2} \left(A_k(\mu + 2\beta) - \beta A_{k+1} - \beta \frac{A_k^2}{A_{k+1}} \right) \left\| z_{k+1} - z_k \right\|^2 \\ & - \frac{\mu}{2} A_k \left(1 - \frac{A_k}{A_{k+1}} \right) \left\| z_{k+1} - y_k \right\|^2. \end{aligned}$$

Substracting $\langle d_{k+1},\, z_{k+1}\rangle$ on both sides and adding/substracting $\langle d_k,\, z_{k+1}\rangle$ and $\langle d_k,\, z_k\rangle$ on the right-hand side, we get

$$\begin{aligned}
A_{k+1}F(y_{k+1}) - c_{k+1} - A_{k+1}H(z_{k+1}) - &\langle d_{k+1},\, z_{k+1}\rangle \\
&\leq A_kF(y_k) - c_k - A_kH(z_k) - \langle d_k,\, z_k\rangle \\
&\quad + \langle d_k,\, z_k\rangle - \langle d_{k+1},\, z_{k+1}\rangle - \langle d_k,\, z_{k+1}\rangle + \langle d_k,\, z_{k+1}\rangle \\
&\quad + (A_{k+1} - A_k)\langle \nabla G(v_k),\, z_{k+1}\rangle - A_k\langle g_H(z_k),\, z_{k+1} - z_k\rangle \\
&\quad - \frac{1}{2}\left(A_k(\mu + 2\beta) - \beta A_{k+1} - \beta\frac{A_k^2}{A_{k+1}}\right)\|z_{k+1} - z_k\|^2 \\
&\quad - \frac{\mu}{2}A_k\left(1 - \frac{A_k}{A_{k+1}}\right)\|z_{k+1} - y_k\|^2 \\
&= A_kF(y_k) - c_k - A_kH(z_k) - \langle d_k,\, z_k\rangle \\
&\quad + \langle d_k - d_{k+1},\, z_{k+1}\rangle \\
&\quad + (A_{k+1} - A_k)\langle \nabla G(v_k),\, z_{k+1}\rangle - \langle A_k g_H(z_k) + d_k,\, z_{k+1} - z_k\rangle \\
&\quad - \frac{1}{2}\left(A_k(\mu + 2\beta) - \beta A_{k+1} - \beta\frac{A_k^2}{A_{k+1}}\right)\|z_{k+1} - z_k\|^2 \\
&\quad - \frac{\mu}{2}A_k\left(1 - \frac{A_k}{A_{k+1}}\right)\|z_{k+1} - y_k\|^2 .
\end{aligned}$$

Now, by first-order optimality conditions of (13), $0 \in d_k + A_k\partial H(z_k) + \beta\partial w(z_k)$. Therefore there exist subgradients $g'_H(z_k) \in \partial H(z_k)$ and $g'_w(z_k) \in \partial w(z_k)$ such that $d_k + A_k g'_H(z_k) = -\beta g'_w(z_k)$. Since the above inequality is true for any $g_H(z_k) \in \partial H(z_k)$, it is in particular true for $g'_H(z_k)$, and thus we have

$$\begin{aligned}
A_{k+1}F(y_{k+1}) - c_{k+1} - A_{k+1}H(z_{k+1}) - &\langle d_{k+1},\, z_{k+1}\rangle \\
&\leq A_kF(y_k) - c_k - A_kH(z_k) - \langle d_k,\, z_k\rangle \\
&\quad + \langle d_k - d_{k+1},\, z_{k+1}\rangle \\
&\quad + (A_{k+1} - A_k)\langle \nabla G(v_k),\, z_{k+1}\rangle + \beta\langle g'_w(z_k),\, z_{k+1} - z_k\rangle \\
&\quad - \frac{1}{2}\left(A_k(\mu + 2\beta) - \beta A_{k+1} - \beta\frac{A_k^2}{A_{k+1}}\right)\|z_{k+1} - z_k\|^2 \\
&\quad - \frac{\mu}{2}A_k\left(1 - \frac{A_k}{A_{k+1}}\right)\|z_{k+1} - y_k\|^2 .
\end{aligned}$$

Since $d_k - d_{k+1} = -(A_{k+1} - A_k)g_k$ we get

$$\begin{aligned}
A_{k+1}F(y_{k+1}) - c_{k+1} - A_{k+1}H(z_{k+1}) - \langle d_{k+1},\, z_{k+1}\rangle \leq\; & A_kF(y_k) - c_k - A_kH(z_k) - \langle d_k,\, z_k\rangle \\
& + (A_{k+1} - A_k)\langle \nabla G(v_k) - g_k,\, z_{k+1}\rangle \\
& + \beta\langle g'_w(z_k),\, z_{k+1} - z_k\rangle \\
& - \frac{1}{2}\left(A_k(\mu + 2\beta) - \beta A_{k+1} - \beta\frac{A_k^2}{A_{k+1}}\right)\|z_{k+1} - z_k\|^2 \\
& - \frac{\mu}{2}A_k\left(1 - \frac{A_k}{A_{k+1}}\right)\|z_{k+1} - y_k\|^2 .
\end{aligned}$$

Finally, by strong convexity of w we have

$$\beta\langle g'_w(z_k),\, z_{k+1} - z_k\rangle \leq \beta\left(w(z_{k+1}) - w(z_k) - \frac{\nu}{2}\|z_{k+1} - z_k\|^2\right),$$

and thus the previous inequality becomes

$$\begin{aligned}
A_{k+1}F(y_{k+1}) - m_{k+1}(z_{k+1}) \leq\; & A_kF(y_k) - m_k(z_k) \\
& + (A_{k+1} - A_k)\langle \nabla G(v_k) - g_k,\, z_{k+1}\rangle \\
& - \frac{1}{2}\left(A_k(\mu + 2\beta) + \beta(\nu - A_{k+1}) - \beta\frac{A_k^2}{A_{k+1}}\right)\|z_{k+1} - z_k\|^2 \\
& - \frac{\mu}{2}A_k\left(1 - \frac{A_k}{A_{k+1}}\right)\|z_{k+1} - y_k\|^2 .
\end{aligned}$$

□

Those three lemmas allow us to prove Proposition 3.

Proof of Proposition 3.

We can rewrite the previous result as

$$\begin{aligned}A_{k+1}F(y_{k+1}) - m_{k+1}(z_{k+1}) \leq\ & A_kF(y_k) - m_k(z_k)\\ &+ (A_{k+1}-A_k)\langle \nabla G(v_k) - g_k,\, z_{k+1} - z_k\rangle\\ &+ (A_{k+1}-A_k)\langle \nabla G(v_k) - g_k,\, z_k\rangle\\ &- \frac{1}{2}\left(A_k(\mu+2\beta) + \beta(\nu - A_{k+1}) - \beta\frac{A_k^2}{A_{k+1}}\right)\|z_{k+1}-z_k\|^2\\ &- \frac{\mu}{2}A_k\left(1-\frac{A_k}{A_{k+1}}\right)\|z_{k+1}-y_k\|^2.\end{aligned}$$

Taking expectation at iteration k conditioned on the previous iterations, we have $\mathbb{E}_k[(A_{k+1} - A_k)\langle \nabla G(v_k) - g_k,\, z_k\rangle = 0]$.

Moreover, by Fenchel-Young inequality we have that for any $\rho > 0$,

$$\langle \nabla G(v_k) - g_k,\, z_{k+1} - z_k\rangle \leq \frac{1}{2\rho}\|\nabla G(v_k) - g_k\|_*^2 + \frac{\rho}{2}\|z_{k+1}-z_k\|^2.$$

Taking expectation and using Assumption 1,

$$(A_{k+1}-A_k)\mathbb{E}_k[\langle \nabla G(v_k) - g_k,\, z_{k+1}-z_k\rangle] \leq \frac{1}{2}(A_{k+1}-A_k)\frac{\sigma^2}{\rho} + \frac{1}{2}(A_{k+1}-A_k)\rho\mathbb{E}_k\|z_{k+1}-z_k\|^2$$

Thus we have

$$\begin{aligned}\mathbb{E}_k\left[A_{k+1}F(y_{k+1}) - m_{k+1}(z_{k+1})\right] \leq\ & A_kF(y_k) - m_k(z_k) + \frac{1}{2}(A_{k+1}-A_k)\frac{\sigma^2}{\rho}\\ &- \frac{1}{2}\left(A_k(\mu+2\beta) + \beta(\nu - A_{k+1}) - \beta\frac{A_k^2}{A_{k+1}} - \rho(A_{k+1}-A_k)\right)\mathbb{E}_k\|z_{k+1}-z_k\|^2\\ &- \frac{\mu}{2}A_k\left(1-\frac{A_k}{A_{k+1}}\right)\mathbb{E}_k\|z_{k+1}-y_k\|^2.\end{aligned}$$

Now, observe that since $0 \leq A_k/A_{k+1} \leq 1$, the term in $\mathbb{E}_k\|z_{k+1}-y_k\|^2$ is non-positive. Therefore, to obtain the final result, it suffices to set A_{k+1} so that the term in $\mathbb{E}_k\|z_{k+1}-z_k\|^2$ cancels out. In other words, we require

$$\begin{aligned}&A_k(\mu+2\beta) + \beta(\nu - A_{k+1}) - \beta\frac{A_k^2}{A_{k+1}} - \rho(A_{k+1}-A_k) = 0\\ \iff\ & A_{k+1}(\beta+\rho) - A_k(\mu+2\beta+\rho) - \beta\nu + \beta\frac{A_k^2}{A_{k+1}} = 0\\ \iff\ & A_{k+1}^2(\beta+\rho) - A_{k+1}\left(A_k(\mu+2\beta+\rho) + \beta\nu\right) + \beta A_k^2 = 0\\ \Leftarrow\ \quad & A_{k+1} = \frac{A_k(\mu+2\beta+\rho) + \beta\nu + \sqrt{\left(A_k(\mu+2\beta+\rho)+\beta\nu\right)^2 - 4(\beta+\rho)\beta A_k^2}}{2(\beta+\rho)}\\ \iff\ & A_{k+1} = \frac{A_k(\mu+2\beta+\rho) + \beta\nu + \sqrt{\beta^2\nu^2 + 2\beta\nu A_k(\mu+2\beta+\rho) + A_k^2\mu^2 + A_k^2\rho^2 + 2A_k^2\mu\rho + 4A_k^2\mu\beta}}{2(\beta+\rho)}\\ \iff\ & A_{k+1} = \frac{A_k(\mu+2\beta+\rho) + \beta\nu + \sqrt{(\beta\nu+\mu A_k)^2 + 4A_k(\beta^2\nu + A_k\mu\beta) + 2\beta\nu A_k\rho + A_k^2\rho^2 + 2A_k^2\mu\rho}}{2(\beta+\rho)}.\end{aligned}$$

Setting $\rho = \sqrt{\mu\beta}$ yields the update for A_{k+1} in Algorithm 1 and proves the proposition.

A.2 Proof of Theorem 1

Proof. Unrolling the recursion in Proposition 3 and taking total expectation, we have

$$\mathbb{E}\left[A_kF(y_k) - m_k(z_k)\right] \leq A_0F(y_0) - m_0(z_0) + A_k\frac{1}{2\sqrt{\mu\beta}}\sigma^2 = -\beta w(z_0) + A_k\frac{1}{2\sqrt{\mu\beta}}\sigma^2$$

Now,

$$m_k(y_\star) = A_k H(y_\star) + \beta w(y_\star) + \sum_{t=0}^{k-1} (A_{t+1} - A_t)\left(G(v_t) - \langle \nabla G(v_t),\, v_t\rangle + \langle g_t,\, y_\star\rangle\right)$$

Taking total expectation on the g_t we get

$$\begin{aligned}\mathbb{E}[m_k(y_\star)] &= A_k H(y_\star) + \beta w(y_\star) + \sum_{t=0}^{k-1} (A_{t+1} - A_t)\,\mathbb{E}\left[G(v_t) + \langle \nabla G(v_t),\, y_\star - v_t\rangle\right] \\ &\leq A_k H(y_\star) + \beta w(y_\star) + A_k G(y_\star) = A_k F(y_\star) + \beta w(y_\star)\end{aligned}$$

where the inequality is by convexity of G. Moreover, $m_k(z_k) \leq m_k(y_\star)$ by construction and thus

$$\begin{aligned}\mathbb{E}[A_k F(y_k) - A_k F(y_\star) - \beta w(y_\star)] &\leq \mathbb{E}[A_k F(y_k) - m_k(y_\star)] \\ &\leq \mathbb{E}[A_k F(y_k) - m_k(z_k)] \\ &\leq -\beta w(z_0) + \frac{A_k}{2\sqrt{\mu\beta}}\sigma^2\end{aligned}$$

and thus

$$\begin{aligned}\mathbb{E}[F(y_k) - F(y_\star)] &\leq \frac{\beta(w(y_\star) - w(z_0))}{A_k} + \frac{\sigma^2}{2\sqrt{\mu\beta}} \\ &= \frac{\beta D_w(y_\star, y_0)}{A_k} + \frac{\sigma^2}{2\sqrt{\mu\beta}}\end{aligned}$$

as $y_0 = z_0$ and $w(y_\star) - w(y_0)$ is equal to $D_w(y_\star, y_0) := w(y_\star) - w(y_0) - \langle g_w(y_0); y_\star - y_0\rangle$ (with $g_w(y_0) \in \partial w(y_0)$) through the choice $g_w(y_0) = 0 \in \partial w(y_0)$, which is valid as y_0 minimizes $w(\cdot)$.

Finally, we can bound A_{k+1} as

$$\begin{aligned}A_{k+1} &\geq A_k \frac{\mu + 2\beta + \sqrt{\mu\beta} + \sqrt{\mu^2 + 4\mu\beta + \sqrt{\mu\beta}^2 + 2\mu\sqrt{\mu\beta}}}{2(\beta + \sqrt{\mu\beta})} \\ &\geq A_k \frac{\mu + 2\beta + \sqrt{\mu\beta} + 2\sqrt{\mu\beta}}{2(\beta + \sqrt{\mu\beta})} \\ &\geq A_k \left(1 + \frac{\sqrt{\mu\beta}}{2(\beta + \sqrt{\beta\mu})}\right) \\ &= A_k \left(1 + \frac{\sqrt{\mu}}{2(\sqrt{\beta} + \sqrt{\mu})}\right)\end{aligned}$$

and thus

$$\mathbb{E}[F(y_k) - F^*] \leq \exp\left(-\frac{k\sqrt{\mu}}{2\left(\sqrt{\beta} + \sqrt{\mu}\right)}\right)\beta D_w(y_\star, y_0) + \frac{\sigma^2}{2\sqrt{\mu\beta}}.$$

□

B Proofs for Accelerated Frank-Wolfe

B.1 Proof of Theorem 2

B.1.1 Proof of feasibility

First we show that $x_k \in K$ for all $k \in \mathbb{N}$.

Proof. We prove this by induction. By assumption $x_0 \in K$. Now suppose $x_k \in K$ for some $k \in \mathbb{N}$. We then have

$$\begin{aligned}
x_{k+1} &= \frac{\beta}{A_{k+1}+\beta}x_0 - \frac{d_{k+1}}{A_{k+1}+\beta} \\
&= \frac{\beta}{A_{k+1}+\beta}x_0 - \frac{d_k + (A_{k+1}-A_k)g_k}{A_{k+1}+\beta} \\
&= \frac{(A_k+\beta)\left(\frac{\beta}{A_k+\beta}x_0 - \frac{d_k}{A_k+\beta}\right) - (A_{k+1}-A_k)g_k}{A_{k+1}+\beta} \\
&= \frac{(A_k+\beta)}{A_{k+1}+\beta}x_k + \frac{A_{k+1}-A_k}{A_{k+1}+\beta}(-g_k) \\
&= \frac{A_k+\beta}{A_{k+1}+\beta}x_k + \left(1 - \frac{A_k+\beta}{A_{k+1}+\beta}\right)(-g_k).
\end{aligned}$$

By induction hypothesis, $x_k \in K$. We also have $-g_k \in K$ since

$$\begin{aligned}
-g_k &= -\frac{1}{m}\sum_{i=1}^{m} g_{k,i} \\
&= \frac{1}{m}\sum_{i=1}^{m} \arg\max_{u \in K} \langle u,\, -v_k + \alpha\Delta_i \rangle.
\end{aligned}$$

In other words, $-g_k$ is a convex combination of elements of K and is thus in K. Therefore $x_{k+1} \in K$ as a convex combination of elements of K. □

B.1.2 Proof of dual gap convergence

Proof. With $H(y) = f^*(y)$, $G(y) = s_\alpha(-y)$, $\beta = \frac{R_K M}{\alpha}$ and $\mu = \frac{1}{L}$, we can apply Proposition 3 to $F = H + G$ and get

$$\mathbb{E}_k[A_{k+1}F(y_{k+1}) - m_{k+1}(z_{k+1})] \leq A_k F(y_k) - m_k(z_k) + (A_{k+1}-A_k)\frac{\sigma^2}{2\sqrt{\mu\beta}}$$

where $m_k(y) = \langle d_k,\, y\rangle + c_k + A_k H(y) + \beta w(y)$, and $z_k = \nabla f(x_k)$. Unrolling the recursion as before and taking total expectation we have

$$\mathbb{E}[A_k F(y_k) - m_k(z_k)] \leq -\beta w(z_0) + A_k \frac{\sigma^2}{2\sqrt{\mu\beta}}.$$

Recall that we set $w(y) = f^*(y) - \langle x_0,\, y\rangle$. Plugging in the function $H = f^*$, and recalling that z_k minimizes $m_k(y)$, we get

$$\begin{aligned}
m_k(z_k) &= \inf_y \{\langle d_k,\, y\rangle + c_k + (A_k+\beta)f^*(y) - \beta\langle x_0,\, y\rangle\} \\
&= c_k - \sup_y \{\langle -d_k + \beta x_0,\, y\rangle - (A_k+\beta)f^*(y)\} \\
&= c_k - (A_k+\beta)f\left(\frac{-d_k+\beta x_0}{A_k+\beta}\right) \\
&= c_k - (A_k+\beta)f(x_k).
\end{aligned}$$

Thus we can conclude that

$$\mathbb{E}[A_k F(y_k) + (A_k+\beta)f(x_k)] \leq -\beta w(z_0) + A_k\frac{\sigma^2}{2\sqrt{\mu\beta}} + c_k,$$

and in particular

$$\mathbb{E}[A_k F(y_k) + A_k f(x_k)] \leq -\beta f(x_\star) - \beta w(z_0) + A_k\frac{\sigma^2}{2\sqrt{\mu\beta}} + c_k. \tag{28}$$

Now,

$$c_k = \sum_{i=0}^{k-1} (A_{i+1} - A_i)(G(v_i) - \langle \nabla G(v_i),\, v_i \rangle).$$

For any $v \in \mathbf{V}$, $G(v) = s_\alpha(-v)$ and thus

$$G(v) - \langle \nabla G(v),\, v \rangle = s_\alpha(-v) + \langle \nabla s_\alpha(-v),\, v \rangle.$$

From Fenchel-Young, $\langle \nabla s_\alpha(-v),\, -v \rangle = s_\alpha(-v) + s_\alpha^*(\nabla s_\alpha(-v))$, and thus

$$G(v) - \langle \nabla G(v),\, v \rangle = -s_\alpha^*(\nabla s_\alpha(-v)).$$

Now, for all $u \in \mathbf{V}^*$,

$$\begin{aligned} s_\alpha^*(u) &= \sup_{v \in \mathbf{V}} \{\langle u,\, v \rangle - s_\alpha(v)\} \\ &\geq \sup_{v \in \mathbf{V}} \{\langle u,\, v \rangle - s(v) - \alpha s_1(0)\} \\ &= s^*(u) - \alpha s_1(0) \\ &= I_K(u) - \alpha s_1(0) \end{aligned}$$

where the inequality is from Proposition 2. In particular, since $\nabla s_\alpha(-v)$ is always feasible, we have $s_\alpha^*(\nabla s_\alpha(-v)) \geq -\alpha s_1(0)$. and thus

$$c_k \leq \sum_{i=0}^{k-1} (A_{i+1} - A_i)\alpha s_1(0) = A_k \alpha s_1(0).$$

We can then rewrite (28) as

$$\mathbb{E}[A_k f^*(y_k) + A_k s_\alpha(-y_k) + A_k f(x_k)] \leq -\beta w(z_0) - \beta f(x_\star) + A_k \frac{\sigma^2}{2\sqrt{\mu\beta}} + A_k \alpha s_1(0).$$

Now,

$$w(z_0) = f^*(z_0) - \langle x_0,\, z_0 \rangle = -f(x_0)$$

by Fenchel-Young and since $z_0 = \nabla f(x_0)$. Thus

$$\mathbb{E}[A_k f^*(y_k) + A_k s_\alpha(-y_k) + A_k f(x_k)] \leq \beta(f(x_0) - f(x_\star)) + A_k \frac{\sigma^2}{2\sqrt{\mu\beta}} + A_k \alpha s_1(0).$$

Finally, from Proposition 2,

$$s_\alpha(-y) \geq s(-y)$$

for any y. We can then conclude

$$\mathbb{E}\left[f^*(y_k) + s(-y_k) + f(x_k)\right] \leq \frac{\beta(f(x_0) - f(x_\star))}{A_k} + \frac{\sigma^2}{2\sqrt{\mu\beta}} + \alpha s_1(0)$$

Bounding A_k as in Theorem 1 yields

$$\mathbb{E}\left[f^*(y_k) + s(-y_k) + f(x_k)\right] \leq \exp\left(-\frac{\sqrt{\mu}}{2(\sqrt{\beta} + \sqrt{\mu})}\right) \beta(f(x_0) - f(x_\star)) + \frac{\sigma^2}{2\sqrt{\mu\beta}} + \alpha s_1(0) \tag{29}$$

It remains to bound the variance σ^2. Using Assumption 2 we have

$$\sigma^2 = \mathbb{E}\left\|g_k - \nabla G(v_k)\right\|_*^2 \leq \frac{4R_K^2 \rho_{\|\cdot\|_*}}{m}.$$

Plugging this back into equation (29), and plugging in the values of $\beta = \frac{R_K M}{\alpha}$ and $\mu = \frac{1}{L}$ gives

$$\mathbb{E}\left[f(x_k) - d(y_k)\right] \leq \exp\left(-k \frac{\sqrt{\frac{1}{L}}}{2\left(\sqrt{\frac{R_K M}{\alpha}} + \sqrt{\frac{1}{L}}\right)}\right) \frac{R_K M}{\alpha}\left(f(x_0) - f(x_\star)\right) + \frac{2R_K^2 \rho_{\|\cdot\|_*}}{m} \sqrt{\frac{\alpha L}{R_K M}} + \alpha s_1(0).$$

Assuming $\frac{R_K M}{\alpha} \geq \frac{1}{L}$ yields the result. □

B.2 Proof of Theorem 3

Proof. Setting

$$\alpha = \min\left\{\frac{\epsilon}{3s_1(0)}, \frac{M\epsilon^2 m^2}{36LR_K^3\rho^2_{\|\cdot\|_*}}\right\},$$

it is easy to verify that

$$\alpha s_1(0) \leq \frac{\epsilon}{3},$$

and that

$$\frac{2R_K^2\rho_{\|\cdot\|_*}}{m}\sqrt{\frac{\alpha L}{R_K M}} \leq \frac{\epsilon}{3}.$$

It remains to compute k such that the first term in the bound of Theorem 2 is also smaller than $\epsilon/3$. This gives

$$\begin{aligned}
&\exp\left(-k\frac{\sqrt{\alpha}}{4\sqrt{LR_K M}}\right)\frac{R_K M}{\alpha}\left(f(x_0) - f(x_\star)\right) \leq \frac{\epsilon}{3} \\
\iff k &\geq \frac{4\sqrt{LR_K M}}{\sqrt{\alpha}}\log\left(\frac{3R_K M(f(x_0) - f(x_\star))}{\epsilon\alpha}\right).
\end{aligned}$$

The $\tilde{O}$-complexity follows directly from plugging the value of α in the bound. □

B.3 Dependence on the norms

In this section, we compute the value of $\rho_{\|\cdot\|_*}$ for different ℓ_p norms when the underlying vector space has dimension d.

B.3.1 Euclidean norm

In the case of the Euclidean norm, we have $\|\cdot\| = \|\cdot\|_* = \|\cdot\|_2$ and thus

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^m g_{k,i} - \nabla G(v_k)\right\|^2 &= \mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^m g_{k,i} + \nabla s_\alpha(-v_k)\right\|^2 \\
&= \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}\left\|g_{k,i} + \nabla s_\alpha(-v_k)\right\|^2 \\
&\quad + \frac{2}{m^2}\sum_{1\leq i<j\leq m}\mathbb{E}[\langle g_{k,i} + \nabla s_\alpha(-v_k),\, g_{k,j} + \nabla s_\alpha(-v_k)\rangle] \\
&= \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}\left\|g_{k,i} + \nabla s_\alpha(-v_k)\right\|^2 \\
&= \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}\left\|-\arg\max_{u\in K}\langle u,\, -v_k + \alpha\Delta_i\rangle + \nabla s_\alpha(-v_k)\right\|^2
\end{aligned}$$

where the second equality simply comes from the properties of the Euclidean norm, and the third one comes from the fact that $g_{k,i} + \nabla s_\alpha(-v_k)$ and $g_{k,j} + \nabla s_\alpha(-v_k)$ are zero-mean independent random variables for all $i \neq j$.
Finally, $\arg\max_{u\in K}\langle u,\, v\rangle \in K$ for any v, and similarly $\nabla s_\alpha(v) = \mathbb{E}\left[\arg\max_{u\in K}\langle u,\, v + \alpha\Delta\rangle\right] \in$

K for all v. Therefore we have

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} g_{k,i} - \nabla G(v_k)\right\|^2 &\leq \frac{1}{m^2}\sum_{i=1}^{m}\max_{u,v\in K}\|u-v\|^2 \\ &\leq \frac{1}{m}\max_{u,v\in K}\left(\|u\|+\|v\|\right)^2 \\ &\leq \frac{1}{m}\max_{u,v\in K} 2\|u\|^2 + 2\|v\|^2 \\ &= \frac{4R_K^2}{m}\end{aligned}$$

and we see that in this case $\rho_{\|\cdot\|_2} = 1$.

B.3.2 ℓ_p-norms for $2 \leq p < \infty$

When $\|\cdot\|_* = \|\cdot\|_p$, we have $\|\cdot\| = \|\cdot\|_q$ for $\frac{1}{p} + \frac{1}{q} = 1$. Since $q \in (1, 2]$, from [6] we know that $\frac{1}{2}\|\cdot\|_q^2$ is $(q-1)$-strongly convex with respect to $\|\cdot\|_q$. Therefore $\frac{1}{2}\|\cdot\|_p^2$ is $\frac{1}{q-1}$-smooth with respect to $\|\cdot\|_p$ [29]. We have

$$\frac{1}{q-1} = \frac{1/q}{1-1/q} = \frac{p}{q} = p-1$$

so $\frac{1}{2}\|\cdot\|_p^2$ is $(p-1)$-smooth with respect to $\|\cdot\|_p$. We now closely follow the proof from [30, Lemma 2]. Let $F = \frac{1}{2}\|\cdot\|_p^2$ and let $Z_i = g_{k,i} - \nabla G(v_k)$ so that $\mathbb{E}\|Z_i\|_p^2 \leq 4R_K^2$. Let $S_i = \sum_{j=1}^{i-1} S_j$. By smoothness of F we have

$$F(S_{i-1} + Z_i) \leq F(S_i) + \langle \nabla F(S_{i-1}),\, Z_i\rangle + \frac{p-1}{2}\|Z_i\|_p^2$$

Taking conditional expectation with respect to $Z_1, \ldots, Z_{i-1}$, since $\mathbb{E}[Z_i] = 0$ we have

$$\begin{aligned}\mathbb{E}_i[F(S_i) \mid Z_1, \ldots, Z_{i-1}] &\leq F(S_{i-1}) + \frac{p-1}{2}\mathbb{E}\left[\|Z_i\|_p^2 \mid Z_1, \ldots, Z_{i-1}\right] \\ &\leq F(S_{i-1}) + \frac{p-1}{2}4R_K^2.\end{aligned}$$

Thus, $F(S_i) - \frac{i(p-1)}{2}$ is a supermartingale and therefore

$$\mathbb{E}[F(S_n)] = \mathbb{E}\left[\frac{1}{2}\left\|\sum_{i=1}^{m} g_{k,i} - \nabla G(v_k)\right\|_p^2\right] \leq \frac{m(p-1)4R_K^2}{2}$$

which shows that $\rho_{\|\|_p} = p-1$.

B.3.3 ℓ_p-norms for $1 \leq p < 2$

Recall that for $\infty \geq q > r \geq 1$,

$$\|\cdot\|_q \leq \|\cdot\|_r \leq d^{1/r-1/q}\|\cdot\|_q. \tag{30}$$

If the norm of interest is $\|\cdot\|_* = \|\cdot\|_p$ for $1 \leq p < 2$, we have

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} g_{k,i} - \nabla G(v_k)\right\|_*^2 &= \mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} g_{k,i} - \nabla G(v_k)\right\|_p^2 \\ &\leq \left(d^{1/p-1/2}\right)^2 \mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} g_{k,i} - \nabla G(v_k)\right\|_2^2 \\ &\leq d^{(2/p-1)}\frac{1}{m}\max_{u,v\in K}\|u-v\|_2^2\end{aligned}$$

where the second inequality comes from the derivation for the Euclidean norm in Appendix B.3.1. Using (30) we get

$$\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} g_{k,i} - \nabla G(v_k)\right\|_*^2 \leq d^{(2/p-1)}\frac{1}{m}\max_{u,v\in K}\|u-v\|_p^2$$
$$= d^{(2/p-1)}\frac{4R_K^2}{m}$$

and thus $\rho_{\|\cdot\|_p} = d^{(2/p-1)}$.

B.3.4 ℓ_∞-norm

When $\|\cdot\|_* = \|\cdot\|_\infty$, we use inequality (30) to get that for any $1 \leq r < \infty$,

$$\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} g_{k,i} - \nabla G(v_k)\right\|_*^2 = \mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} g_{k,i} - \nabla G(v_k)\right\|_\infty^2$$
$$\leq \mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} g_{k,i} - \nabla G(v_k)\right\|_r^2.$$

Now, if $r \geq 2$, from Appendix B.3.2 we have that

$$\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} g_{k,i} - \nabla G(v_k)\right\|_r^2 \leq \frac{r-1}{m}\max_{u,v\in K}\|u-v\|_r^2$$
$$\leq \frac{(r-1)}{m}d^{2/r}\max_{u,v\in K}\|u-v\|_\infty^2$$
$$= \frac{(r-1)4R_K^2}{m}d^{2/r}.$$

Taking $r = 2 + \log d$ (so that $r \geq 2$), we then have

$$\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} g_{k,i} - \nabla G(v_k)\right\|_*^2 \leq \frac{(r-1)4R_K^2}{m}e^{\frac{2}{r}\log(d)}$$
$$= \frac{4R_K^2}{m}(\log d + 1)e^{\frac{2\log d}{2+\log d}}$$
$$\leq \frac{4R_K^2}{m}2(\log d + 1)$$

and thus we have $\rho_{\|\cdot\|_\infty} = 2(\log d + 1)$.

B.3.5 General norm

For a general norm, since we are in a finite-dimensional space, there exist constants $c, C > 0$ such that $c\|\cdot\|_2 \leq \|\cdot\|_* \leq C\|\cdot\|_2$. We then have

$$\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} g_{k,i} - \nabla G(v_k)\right\|_*^2 \leq C^2\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} g_{k,i} - \nabla G(v_k)\right\|_2^2$$
$$\leq C^2\frac{1}{m^2}\sum_{i=1}^{m}\max_{u,v\in K}\|u-v\|_2^2$$
$$\leq \frac{C^2}{c^2}\frac{1}{m^2}\sum_{i=1}^{m}\max_{u,v\in K}\|u-v\|_*^2$$
$$\leq \frac{C^2}{c^2}\frac{4R_K^2}{m}$$

where the second inequality comes from the derivation in Appendix B.3.1. Thus we $\rho_{\|\cdot\|_*} \leq \frac{C^2}{c^2}$.

C Experimental Details

In the experiments we only consider Euclidean norms so that $\|\cdot\| = \|\cdot\|_*$.
If the entries of z are independently distributed according to a Gumbel distribution with location 0 and scale 1, the probability density function reads

$$p(z) = e^{-\left(\sum_{i=1}^d z_i + e^{-z_i}\right)}$$

so that $\eta(z) = \sum_{i=1}^d z_i + e^{-z_i}$. Thus we have

$$\nabla \eta(z) = \begin{pmatrix} 1 - e^{-z_1} \\ 1 - e^{-z_2} \\ \vdots \\ 1 - e^{-z_d} \end{pmatrix}$$

and

$$\|\nabla \eta(z)\|^2 = \sum_{j=1}^d (1 - e^{-z_j})^2$$

Algorithm 2 and Algorithm 3 require a bound on the value of M. We compute it now.

$$\begin{aligned} M^2 &= \mathbb{E}\,\|\nabla \eta(Z)\|^2 \\ &= \int_{\mathbb{R}^d} \sum_{j=1}^d (1 - e^{-z_j})^2 e^{-\left(\sum_{i=1}^d z_i + e^{-z_i}\right)} dz \\ &= d \int_{\mathbb{R}^d} (1 - e^{-z_1})^2 e^{-\left(\sum_{i=1}^d z_i + e^{-z_i}\right)} dz \\ &= d \int_{\mathbb{R}} (1 - e^{-z_1})^2 e^{-(z_1 + e^{-z_1})} \int_{\mathbb{R}} e^{-(z_2 + e^{-z_2})} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} e^{-(z_d + e^{-z_d})} dz_d \ldots dz_2 dz_1 \end{aligned}$$

For any $i \geq 2$,

$$\int_{\mathbb{R}} e^{-(z_i + e^{-z_i})} dz_i = 1.$$

Moreover, one can check that

$$\int (1 - e^{-z_1})^2 e^{-(z_1 + e^{-z_1})} dz_1 = e^{-e^{-z_1}} + e^{-2z_1 - e^{-z_1}} + C$$

where C is some constant. Taking limits one gets

$$\int_{\mathbb{R}} (1 - e^{-z_1})^2 e^{-(z_1 + e^{-z_1})} dz_1 = 1$$

and thus

$$M^2 = d.$$

The derivation in the case of a multivariate normal distribution is similar.